GeneSpring
User Manual

Version 7.0

© 2004 Silicon Genetics. All rights reserved. GeneSpring, GeneSpider, GenEx, Signet, MetaMine, ScriptEditor and MicroSift are trademarks of Silicon Genetics. All other products, including but not limited to Affymetrix GeneChip®, Affymetrix Global Scaling™, GenBank, Microsoft Excel®, Microsoft Notepad®, Pico™, SimpleText© and Adobe FrameMaker®, are the trademarks of their respective holders.

Table of Contents

Table of Contents . . . . . Index-iii

About This Guide xvii

1 Introduction 1-1

What is GeneSpring? 1-1
Identify Targets Reliably and Quickly 1-1
Uncover Statistically Meaningful Results 1-2
Displaying Expression Data 1-2
Predict Clinical Outcomes 1-2
Characterize Novel Expression Patterns 1-3
Test Complex Hypotheses 1-3
Ensure Data Quality 1-3
Record and Automate Complex Analysis Tasks 1-4
Import and Compare Expression Data from Multiple Platforms 1-4
Export Data and Images in Standard Formats 1-4
Take Advantage of World-class Technical Support 1-4
GeneSpring Key Features 1-5
Advanced Statistical Tools 1-5
Data Clustering 1-5
Visual Filtering 1-5
3D Data Visualization 1-5
Data Normalization 1-6
Pathway Views 1-6
Search for Similar Samples 1-6
Support for MIAME Compliance 1-6
Scripting 1-6
PCA on Conditions 1-6
Organism-Specific KEGG and GenMAPP Pathways Import 1-7
What's New in This Release? 1-7
Genome Import Wizard 1-7
Project-based Navigation 1-8
K-Nearest Neighbors Class Predictor 1-8
Support Vector Machines 1-8
Master Table of Genes Editor 1-8
MAGE-ML Export 1-9
Basic Script Filters 1-9
Find Significant Parameters 1-9
GeneSpring Java API 1-9

Data Object Modification in Signet. 1-10

2 GeneSpring Interface 2-1

GeneSpring Window 2-1
Menu Bar 2-2
File Menu 2-2
Edit Menu 2-4
View Menu 2-5
Experiments Menu 2-7
Colorbar Menu 2-7
Filtering Menu 2-8
Tools Menu 2-8
Annotations Menu 2-9
Window Menu 2-10
Help Menu 2-10
Navigator 2-11
Folders 2-12
Menus 2-14
Special Menus 2-15
Genome Browser 2-16
Menus 2-17
Special Menus 2-17
Button Bar 2-18
Colorbar 2-19
Image Window 2-20
Inspectors 2-21
Wizards 2-22

3 Getting Started 3-1

Requirements 3-1
Windows 3-1
Macintosh 3-1
UNIX 3-2
Installing GeneSpring 3-2
Installing GeneSpring from CD 3-2
Installing from the Web 3-2
Starting GeneSpring 3-2
Obtaining a License Key 3-3
Setting Memory Usage Options 3-3
Verifying Virtual Memory 3-3
Working with Previous Releases of GeneSpring 3-4
Updating GeneSpring 3-4
Learning to use GeneSpring 3-4
GeneSpring Basics 3-5
Workflow 3-5
Terminology 3-6
Basic Actions 3-8

Basic Functions 3-13
Setting Preferences. 3-15
Opening the Preferences Window 3-15
Data Files Tab 3-16
Database Tab 3-17
Colors Tab 3-17
Browser Tab 3-20
Firewall Tab 3-20
System Tab 3-21
Signet Tab 3-22
Computation Tab 3-23
Miscellaneous Tab. 3-24

4 Importing Genomes 4-1

Overview 4-1
Import Options 4-2
Import New Genome Wizard 4-2
Importing Genomes from Silicon Genetics 4-2
Pre-made Genomes 4-2
Importing Genomes from Silicon Genetics. 4-2
Importing Genomes from GenBank/EMBL Files 4-4
Importing Tab-delimited Genome Files. 4-7
Selecting Annotation Files 4-7
Importing Genome Sequences. 4-12
Managing Web Links. 4-15
Standard Gene URLs. 4-16
Managing Gene Links 4-16
Managing Experiment Links 4-22
Saving New Genomes 4-27
Workflow for Saving New Genomes 4-27
Naming New Genomes 4-27
Updating Annotations with GeneSpider 4-29
Making Gene Lists from Annotations 4-30
Building Ontologies 4-31
Importing KEGG Pathways 4-32
Building Homology Tables 4-34
Obtaining Supplemental Information 4-36

5 Working With Experiments 5-1

Before You Begin 5-1
File Formats. 5-1
Memory Requirements for Loading Experiments. 5-2
Using Data Preprocessors 5-2

Loading Experiments 5-3
Loading Expression Data from Files 5-3
Using the Column Editor 5-5
Preprocessing Data Files 5-13
Importing Selected Files 5-14
Handling Ambiguous Gene Identifiers 5-15
Selecting Signal and Control Files 5-16
Selecting Samples with Multiple Files 5-17
Managing Data Files with Extra Genes 5-18
Importing Sample Attributes 5-19
Saving New Experiments 5-23
Creating New Experiments 5-25
Create New Experiment Window 5-25
Creating New Experiments 5-28
Copying and Pasting Experiments 5-30
Preparing to Paste 5-30
Common Mistakes in Pasting 5-34
Pasting an Experiment into GeneSpring 5-34
Copying an Experiment Out of GeneSpring 5-35
Using the Default Normalizations 5-35
Setting Up Experiment Parameters 5-37
Experiment Parameters 5-37
Parameter Values 5-37
Using Parameters 5-37
Parameters and Experiment Interpretations 5-38
Using the Experiment Parameters Window 5-39
Working with Samples 5-45
Sample Manager Window 5-45
Opening the Sample Manager 5-47
Filtering Samples 5-47
Finding Similar Samples 5-53
Viewing Similar Sample Results 5-57
Working with Sample Attributes 5-58
Sample Attributes 5-58
Editing Sample Attributes 5-59
Importing Sample Attributes from Experimental Parameters 5-59
Creating New Sample Attributes 5-60
Copying and Pasting Data in the Edit Attributes Window 5-62
Deleting Sample Attributes 5-62
Using the Standard Attributes Editor 5-63
Setting Up Experiment Interpretations 5-65
Parameters and Experiment Interpretations 5-65
Experiment Interpretation Window 5-66
Vertical Axis Modes 5-67
Parameter Display Settings 5-70
Opening the Experiment Interpretation Window 5-71
Changing Experiment Interpretation 5-72
Finding Experiment Interpretations 5-73

Deleting Experiment Interpretations 5-73
Using Cross-Gene Error Models 5-73
Defining the Cross-Gene Error Model 5-76
Enabling the Cross-Gene Error Model 5-76
Technical Details for the Cross-Gene Error Model. 5-77
References 5-78

6 Managing Genomes 6-1

Overview 6-1
Managing Genomes 6-1
Genome Manager Window 6-2
Opening the Genome Manager 6-4
Signet Warnings 6-4
Importing GeneSpring Zip Files 6-5
Exporting GeneSpring Zip Files 6-6
Adding New Folders 6-6
Deleting Folders 6-7
Renaming Folders 6-7
Renaming Genomes 6-8
Deleting Genomes 6-9
Inspecting Genomes. 6-10
Genome Inspector Window. 6-10
Opening the Genome Inspector. 6-20
Editing Gene Properties. 6-21
Viewing Genome History 6-21
Importing Genomic Sequences 6-22
Building Homology Tables 6-23
Managing Gene Links 6-25
Managing Experiment Links 6-31
Managing Genes and Annotations 6-35
Master Table of Genes 6-36
Master Table of Genes Editor Window. 6-36
Which Annotations are Retrieved? 6-41
Opening the Master Table of Genes Editor. 6-42
Editing Columns 6-42
Creating Columns for EBI Submissions 6-44
Sorting Data. 6-45
Adding Genes 6-45
Deleting Genes 6-46
Importing Genes and Annotations. 6-46
Editing Annotations. 6-53
Finding Genes 6-54
Finding and Replacing Text 6-57
Copying Data to the Clipboard 6-58
Saving Data to a File 6-58

Adding Other Genomic Elements 6-59

7 Finding, Selecting, and Inspecting Data 7-1

Finding Genes 7-1
Using the Genome Browser 7-1
Finding Genes 7-3
Using the Advanced Find Gene Function 7-4
Searching for Data Locally or Remotely 7-6
Selecting Genes 7-14
Selecting a Single Gene 7-14
Using Inspectors 7-15
Using the Gene Inspector 7-16
Using the Sample Inspector 7-19
Using the Experiment Inspector 7-25
Using the Condition Inspector 7-30
Using the Gene List Inspector 7-33
Using the Classification Inspector 7-37
Working with Projects 7-40
Project Functions 7-40
Project Assignment 7-41
Filter Navigator Window 7-41
Show Menu 7-45
Opening the Filter Navigator Window 7-45
Assigning Data Objects to Projects 7-46
Assigning New Experiments to Projects 7-50

8 Viewing Data 8-1

Managing Window Elements 8-1
Linking Windows 8-1
Splitting Windows 8-1
Using Bookmarks to Save Your Place 8-3
Showing or Hiding Window Display Elements 8-5
Displaying Expression Data 8-6
Using the Blocks View 8-7
Using the Graph View 8-9
Using Bar Graph View 8-12
Using Physical Position View 8-15
Using the Scatter Plot View 8-20
Using the 3D Scatter Plot View 8-25
Using the Tree View 8-30
Using the Ordered List View 8-38
Using the Array Layout View 8-40
Using the Pathway View 8-42
Using the Compare Genes to Genes View 8-44
Using the Graph by Genes View 8-46
Using the Spreadsheet View 8-50
Using the Condition Scatter Plot 8-52
Changing Common Display Options 8-57

Opening the Display Options Window 8-57
Changing Display Options for the Legend 8-57
Changing Display Options for Color Settings. 8-60

9 Filtering Data 9-1

Filtering 9-1
Filtering on Gene Lists. 9-1
Gene Filters 9-2
Filtering Menu. 9-2
Filter Window 9-3
Data Types for Restrictions. 9-4
Filtering on Expression Level 9-5
Filtering on Fold Change. 9-7
Filtering on Error. 9-10
Filtering on Confidence. 9-12
Filtering on Parameters 9-16
Filtering on Flags. 9-21
Filtering on Gene List Numbers 9-27
References 9-29
Using Advanced Filters 9-29
Creating Advanced Filters. 9-31
Saving Advanced Filters 9-32
Filtering Data Objects Assigned to Projects 9-33
Filtering Assigned and Unassigned Projects with the Show Menu. 9-33
Filtering Projects Assigned to Data Objects 9-34

10 Normalizing Data. 10-1

Experiment Normalizations 10-1
Normalizing Experiments 10-2
Experiment Normalizations Window 10-2
Assumptions 10-3
Normalization Methods. 10-3
Opening the Experiment Normalizations Window 10-3
Adding Normalization Steps 10-4
Normalizing Specific Samples 10-4
Editing Normalization Steps 10-5
Removing Normalization Steps. 10-6
Reordering Normalization Steps 10-6
Applying Default Normalizations 10-6
Viewing Detailed Descriptions of Normalizations 10-7
Saving a Normalization Scenario 10-8
Working with Saved Scenarios 10-8
Normalization Warnings 10-9
Data Transformations. 10-10

SAGE Transformation. 10-10
Real Time PCR Transformation 10-10
Subtract Background Based on Negative Controls 10-10
Set Measurements less than 0.01 to 0.01. 10-11
Transform from Log to Linear Values 10-11
Dye Swap 10-11
Normalization Methods 10-12
Start with Pre-Normalized Values. 10-12
Per-spot Normalization 10-12
Per-chip Normalizations 10-15
Per-gene Normalizations 10-20
Normalization Strategies for Specific Technologies 10-24
Normalization of Affymetrix Data 10-24
Normalization of Two-color Microarray Data 10-25
Region Normalization 10-25
Handling Repeated Measurements 10-26
Negative Control Strengths 10-28
References 10-29

11 Annotating Genes 11-1

GeneSpider. 11-1
Databases 11-1
Silicon Genetics. 11-2
GenBank 11-2
LocusLink 11-4
UniGene. 11-4
Updating Annotations with GeneSpider 11-4
Limitations on Using GeneSpiders 11-5
Updating Annotations from Silicon Genetics 11-5
Updating Annotations from GenBank, LocusLink, or UniGene. 11-6
GeneSpider Errors Window 11-8
Problems with the Map Location Annotations 11-8
Which Annotations are Retrieved? 11-8
Building Ontologies 11-9
Using Pathways 11-10
KEGG Pathways 11-11
GenMAPP Pathways 11-11
Importing KEGG Pathways. 11-11
Importing JPEG and GIF Pathway Image Files 11-12
Adding Genes to Pathways 11-12
Making Gene Lists from Pathways 11-15
Pathway Commands 11-15

12 Working with Gene Lists 12-1

Gene List 12-1
Managing Gene Lists 12-2
Displaying Gene Lists 12-2
Gene List Editor Window 12-3

Opening the Gene List Editor 12-5
Creating or Editing Gene Lists 12-5
Saving Gene Lists 12-6
Deleting Gene Lists 12-8
Filtering Gene Lists 12-8
Filtering on Gene Lists 12-8
Filtering on Lists 12-8
Filtering on Annotations 12-9
Showing All Genes 12-10
Inspecting Gene Lists 12-11
Gene List Inspector Window 12-11
Opening the Gene List Inspector 12-14
Finding Similar Gene Lists 12-15
Configuring Preferences for Gene List Searches 12-15
Making Gene Lists 12-16
Making Gene Lists from Annotations 12-16
Making Gene Lists from Venn Diagrams 12-18
Making Gene Lists from Classifications 12-19
Making Gene Lists from Selected Genes 12-21
Making Gene Lists from Expression Profiles 12-21
Working with Homology Tables 12-23
Build Homology Tables Window 12-24
Supported Organisms 12-25
Homology Table Examples 12-25
Building Homology Tables 12-26
Viewing Homology Tables 12-27
Deleting Homology Tables 12-28

13 Finding Differentially Expressed Genes 13-1

Overview 13-1
Statistical Analysis (ANOVA) 13-1
One-Way ANOVA 13-2
Two-Way ANOVA 13-2
Statistical Analysis (ANOVA) Window 13-3
Before You Begin 13-5
Performing One-way ANOVA 13-5
Performing One-way ANOVA 13-5
Technical Details for One-way ANOVA 13-9
Multiple Testing Corrections 13-12
Post Hoc Tests 13-13
Performing Two-way ANOVA 13-15
Performing Two-way ANOVA 13-15
Technical Details for Two-Way ANOVA 13-18
Non-Parametric Two-way ANOVA (Friedman’s Test) 13-22
Interpreting ANOVA Results 13-23
ANOVA without Post Hoc Test 13-23
ANOVA with Post Hoc Tests 13-23
Two-way ANOVA 13-25

Viewing Generated P-values 13-27

14 Finding Genes with Similar Expression Profiles 14-1

Finding Potential Regulatory Sequences 14-1
Find Potential Regulatory Sequence Window. 14-2
Finding Regulatory Sequences 14-2
Entering Specific Regulatory Sequences 14-4
Viewing Regulatory Sequence Search Results 14-6
Interpreting Regulatory Sequence Search Results 14-6
Viewing the Conjectured Nucleotide Sequence 14-7
Finding Similar Genes 14-9
Correlation of Expression Change 14-11
Complex Correlations Window. 14-11
Technical Details for Similarity Measures 14-15
Performing Complex Correlations 14-21
Finding Similar Genes Based on Correlations 14-23

15 Clustering and Characterizing Data 15-1

Clustering. 15-1
Performing Clustering Analysis. 15-1
Clustering Window 15-2
K-Means Clustering 15-2
Classifications 15-3
Gene Tree 15-3
Condition Tree. 15-4
Self-Organizing Maps 15-4
QT Clustering 15-5
Performing K-means Clustering 15-5
Performing Gene Tree Clustering 15-8
Performing Condition Tree Clustering 15-11
References for Hierarchical Clustering 15-13
Performing Self-Organizing Map Clustering 15-14
Performing QT Clustering. 15-17
Adding or Removing Experiments 15-19
Experiment Weight 15-19
Similarity Measures. 15-20
Performing Principal Components Analysis 15-21
Running a PCA on Genes 15-21
Viewing PCA on Genes Results 15-22
Running a PCA on Conditions 15-23
Viewing PCA on Conditions Results 15-24
Interpreting PCA Results. 15-26
Performing Class Prediction Analysis 15-29
K-Nearest Neighbors. 15-29
Support Vector Machines Kernel Function. 15-30
Performing a Class Prediction Analysis for K-Nearest Neighbors 15-30
Interpreting Class Prediction Analysis Results 15-32
Technical Details for the Class Prediction 15-33

Performing a Support Vector Machines Analysis 15-34
Finding Significant Parameters 15-37
Finding Significant Parameters Using ANOVA 15-37
Finding Significant Parameters Using an Association Test 15-39
Finding Significant Parameters Using Correlation to Parameters. 15-41
Interpreting Finding Significant Parameters Results 15-43
Technical Details for the Find Significant Parameters 15-45

16 Using Scripts, External Programs, and Plugins 16-1

Working with Scripts 16-1
Scripts Folder 16-2
Pre-defined Scripts 16-2
Configuring Script Preferences 16-3
Running Scripts Locally 16-4
Running Scripts Remotely. 16-9
Viewing the Results of a Remote Job 16-11
Handling Scripts that Generate Long Names or Invalid Characters 16-12
Uploading Scripts to Signet. 16-13
Inspecting Scripts. 16-13
Script Inspector Window. 16-13
Opening the Script Inspector Window 16-14
Editing Script Details 16-14
Using the Script Editor. 16-15
Script Editor Terminology. 16-15
Script Editor Window 16-16
Opening the Script Editor Window 16-20
Working with Building Blocks 16-20
Building Block Types 16-20
Inputs and Outputs. 16-21
Knobs. 16-21
Sample Script 16-21
Pre-defined Building Blocks 16-23
Opening Building Blocks 16-38
Opening Pre-defined Building Blocks 16-38
Using Building Blocks from External Programs. 16-39
Managing Scripts 16-40
Creating Scripts 16-40
Defining Knobs 16-41
Converting Knobs to Input 16-41
Converting Input Back to Knobs. 16-43
Saving a Building Blocks as a Script 16-43
Arranging Inputs and Outputs in the Browser. 16-43
Dynamically Naming Script Outputs 16-43
Creating Useful Notes for Output Objects 16-44
Saving Scripts 16-45
Saving Scripts to Subfolders 16-46
Moving Scripts 16-46

Warning Messages 16-46
Getting Help with Scripts. 16-46
Working with External Programs 16-47
External Program Interface 16-47
External Programs Folder 16-47
External Program Script Building Blocks 16-47
New External Program Window 16-48
Creating External Programs 16-50
Running External Programs 16-56
External Program Example 16-56
Inspecting External Programs 16-58
Working with Plugins 16-63
Plugin Types 16-63
Working with Data Preprocessor Plugins 16-63
Working with Interactive Plugins 16-64
Working with Script Plugins 16-66

17 Exporting Data Files 17-1

Exporting Images 17-1
Saving a Genome Browser Image 17-1
Saving a Colorbar Image 17-3
Saving a Venn Diagram Image 17-3
Saving the Active Window Image. 17-3
Saving the Entire Window Image 17-4
Printing Images 17-4
Exporting Gene Lists 17-5
Dragging Gene Lists out of GeneSpring 17-5
Copying and Pasting Gene Lists 17-6
Exporting Annotated Gene Lists 17-6
Copy Annotated Gene List Window 17-6
Copying Annotated Gene Lists 17-9
Saving Annotated Gene Lists 17-10
Exporting MAGE-ML Data 17-10
MAGE-ML Data for Publication 17-11
CompositeSequence Identifiers 17-11
Affymetrix Probesets. 17-12
Exporting MAGE-ML Data 17-13

18 Using Signet 18-1

Signet 18-1
Configuring GeneSpring to Access Signet 18-2
Adding a New Signet Server 18-3
Editing Settings for a Signet Server 18-4
Deleting a Signet Server 18-4
Logging in to Signet. 18-4
Uploading Data to Signet. 18-5
Bulk Upload of Data to Signet. 18-7

Editing Data in Signet 18-10
Changing Ownership of Data in Signet 18-10

A Custom Databases and GeneSpring A-1

Database A-1
Open Database Connectivity A-1
Structured Query Language A-2
SQL Call Level Interfaces A-2
Genetic Analysis Technology Consortium A-2
Databases and GeneSpring A-3
Adding an Experiment from a Database A-3
Creating a New ODBC Source A-3
Testing Your ODBC Connection A-4
Connecting your Database to GeneSpring A-4
Configuration File Reference A-4
Tag Reference A-5
Tag Definitions A-8
Getting Data from a Database A-22
Using Complex Databases A-22

B Name Restrictions for Data Objects B-1

C Array Layout C-1

Creating an Array Layout C-1
Array Layout Format C-1
Examples for Arrays C-2

Glossary Glossary-1

Index Index-1

About This Guide

The *GeneSpring® User Guide* explains how to use GeneSpring® 7 to visualize, analyze, and interpret genomic expression data. Information in this guide applies the Windows, Solaris, and Macintosh operating systems. This chapter covers the following topics:

- Audience
- Organization
- Conventions
- Product Contents
- Printing the Guide
- Support and Training

Audience

This guide is intended for scientists, students, and academicians in biotechnology, pharmaceutical, and academic organizations who require the use of sophisticated analysis tools to advance their research in mRNA microarray analysis.

It is assumed that the reader understands experimental design, statistical analysis techniques, and data interpretation. It is also assumed that the user has a general knowledge of Microsoft Windows, Solaris, or Macintosh operating systems; understands the desktop interface for these platforms; and knows how to use web-based applications.

Organization

This guide is organized as follows:

- Chapter 1, “Introduction”—Provides an overview of GeneSpring and describes the new features in this release.
- Chapter 2, “GeneSpring Interface”—describes the GeneSpring graphical user interface, including the GeneSpring window, browser, menus, inspectors, and wizards.
- Chapter 3, “Getting Started”—Lists the requirements for using GeneSpring, explains how to install GeneSpring, describes resources for learning GeneSpring, and presents basic procedures.
- Chapter 4, “Importing Genomes”—Explains how to import genomes from Silicon Genetics, import data files from GenBank/EMBL databases, and create custom genomes for use in data analysis.

- Chapter 5, “Working With Experiments”—Explains how to import expression data; create experiments; and set up experiment normalizations, parameters, and interpretations. Also explains how to use cross-gene error models to ensure reliable measurements even when an experiment has a limited number of sample replicates.
- Chapter 6, “Managing Genomes”—Explains how to manage genomes, annotations, and sample data using the Genome Manager, Master Table of Genes Editor, and Sample Manager tools.
- Chapter 7, “Finding, Selecting, and Inspecting Data”—Explains the different options that are available to find, select, and inspect genomes or associated data objects.
- Chapter 8, “Viewing Data”—Explains how to display data using different views.
- Chapter 9, “Filtering Data”—Describes the different filters that are available for finding genes and arrays in expression data.
- Chapter 10, “Normalizing Data”—Describes the different normalizations options that are available to transform data in preparation for analysis. Techniques include intensity-dependent (LOWESS) normalization, global per-chip or per-gene normalization, and normalization to specified control samples.
- Chapter 11, “Annotating Genes”—Explains how to annotate data using GeneSpiders and create Gene Ontology (GO) ontologies.
- Chapter 12, “Working with Gene Lists”—Explains how to create, filter, and use gene lists to mine data for further analysis.
- Chapter 13, “Finding Differentially Expressed Genes”—Explains how to analyze data using one-way and two-way analysis of variance (ANOVA) techniques.
- Chapter 14, “Finding Genes with Similar Expression Profiles”—Explains how to find potential regulatory sequences or use correlation techniques to find genes with similar expression profiles.
- Chapter 15, “Clustering and Characterizing Data”—Explains how to perform a cluster analysis, Principal Components Analysis, and class prediction analysis using k-nearest neighbors or Support Vector Machines techniques.
- Chapter 16, “Using Scripts, External Programs, and Plugins”—Explains how to work with scripts, the Script Inspector, and the Script Editor. Also explains how to use external programs and plugins to perform data manipulation and analysis.
- Chapter 17, “Exporting Data Files”—Explains how to export images, gene lists, and MAGE-ML data from GeneSpring.
- Chapter 18, “Using Signet”—Explains how to use Signet to upload, copy, or delete genomes on a Signet server.
- Appendix A, “Custom Databases and GeneSpring”—Explains how to connect to a database, add experiment data from a database, and enter the database into GeneSpring,
- Appendix B, “Name Restrictions for Data Objects”—Describes naming restrictions for data objects in GeneSpring.

- Appendix C, “Array Layout”—Explains the format for creating an array layout that can be used for data analyses or displayed in GeneSpring views.
- Glossary—Explains terminology used in this guide.

Conventions

The following conventions are used in this guide:

Convention	Description
Bold	Indicates the names of menus, commands, options, and parameters in the GeneSpring window.
`Monospaced font`	Represents a location in a directory or example text.
italics	Represents a file type or book title.
Link	Represents a hyperlink to other topics in this guide or to web sites on the Internet.

Product Contents

The following items are included in the GeneSpring 7 product:

Software

The GeneSpring 7 release includes software for the Windows, Solaris, or Mac operating systems. The software is available for download from the Silicon Genetics web site. Alternatively, you can obtain the *GeneSpring 7* CD-ROM from Silicon Genetics.

Documentation

- *GeneSpring User Guide*—Documentation for using GeneSpring *7* (this guide).
- *GeneSpring API JAVADOC*—Documentation for the GeneSpring JAVA API in JAVADOC form. Provides information on creating data preprocessors, script plugins, and interactive plugins. The JAVADOC is provided in HTML format.
- *GeneSpring Tutorial*—Tutorial to get started with using GeneSpring.
- *Release Notes*—Latest information on new features, product requirements, and known issues.

Printing the Guide

Silicon Genetics documentation is provided in Adobe Portable Document Format (PDF). Adobe PDF files are compact and can be viewed, navigated, and printed using the Adobe Acrobat® Reader® software. The software is available at no charge from the Adobe web site at: http://www.adobe.com.

After installing Adobe Acrobat Reader, you can open the PDF in Adobe Acrobat Reader and print the document in its entirety.

Support and Training

Live technical support for all Silicon Genetics products is available from 7 A.M. - 5 P.M. Pacific Standard Time. Please contact us at techsupport@silicongenetics.com or call us at +1-(866) SIG SOFT(866) 744 7638 (in the US) or +81 (70) 5072-2085 (Japan) or +44 (0) 1259 751 833 (Europe). Support via email is also available at: support@silicongenetics.com.

Workshops

Learn how to better analyze your data using GeneSpring. We offer comprehensive on-site workshops run by experienced technical support scientists that cater to all skill levels. Go to http://www.silicongenetics.com/workshops.html for more information.

FAQs & Tech Notes

FAQs are frequently asked questions for those new to or considering our products. Our tech notes provide answers to common GeneSpring and Signet questions. Go to for more information.

Online Training

Learn to use product features by participating in Webinars, attending workshops, viewing online animations, or consulting analysis guides. Go to http://www.silicongenetics.com/Support/training.html for more information.

Analysis Guides

Our *Analysis Guides* are designed to provide an in-depth look at selected GeneSpring features. Go to http://www.silicongenetics.com/Support/guides.html for more information.

GeneSpring User Group

We provide an email list for GeneSpring users to discuss various issues. Go to http://www.silicongenetics.com/Support/egroup.html for more information.

GeneSpring Extras

GeneSpring extras include information on where to find user manuals, MIAME-compliant standard attribute files, pathway files, the GeneSpring-R integration package, and other GeneSpring resources. Go to http://www.silicongenetics.com/resources.html for more information.

1

Introduction

This chapter introduces you to the GeneSpring 7 microarray data analysis application from Silicon Genetics. It covers the following topics:

- What is GeneSpring?
- GeneSpring Key Features
- What's New in This Release?

What is GeneSpring?

GeneSpring is a powerful visualization and analysis solution designed for use with genomic expression data. It is capable of displaying and analyzing large data sets on standard desktop computers. Developed for use in academic, biotechnology, and pharmaceutical organizations, its flexible statistical tools analyze data from virtually any source.

GeneSpring is widely regarded as the gold standard for expression data analysis. When you use GeneSpring, you join thousands of elite scientists who rely on its sophisticated analysis techniques to advance their research. Analyses conducted with GeneSpring have been cited in over 500 peer-reviewed articles in leading scientific publications around the world. Designed to meet the needs of the individual researcher, GeneSpring seamlessly interfaces with the Silicon Genetics Signet™ software, which provides a highly-scalable platform for enterprise-level genomic research.

Identify Targets Reliably and Quickly

Using GeneSpring's sophisticated filtering tools, you can simply and confidently identify genes that are affected by novel drug treatments or experimental conditions. A variety of intuitive visual interfaces allow even novice users to select genes with specific expression patterns. More sophisticated users can take advantage of the advanced filtering window, which allows users to create complex Boolean expressions to identify genes that meet highly-specific, customizable criteria. Once created, filters can easily be saved to standardize critical laboratory procedures, or can be shared with other researchers via the Signet system.

Uncover Statistically Meaningful Results

GeneSpring's robust statistical tools give researchers flexibility in designing complex experimental structures, and provide confidence in analyzing results. With analysis of variance (ANOVA) techniques, you can identify differentially expressed genes when the groups being compared are defined by one or two parameters. GeneSpring provides in-depth techniques for analyzing and visualizing genes differentially expressed across independent groups. These techniques include:

- One-way and two-way ANOVA
- Multiple Testing Corrections
- False Discovery Rate Prediction
- Tukey and Student-Newman-Keuls post hoc tests

Displaying Expression Data

GeneSpring displays expression data in a manner that will help you conceptualize important results and enhance your publications. With the Pathway viewer, genes and their expression patterns can be visually characterized based on their location within a cellular pathway. GeneSpring can import pathways from numerous sources, including native import from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Data can be viewed with highly-customizable visualization tools including:

- 2D and 3D scatter plots
- 2D dendrograms
- Chromosome maps
- Pathway diagrams
- Venn diagrams
- Classification views

Predict Clinical Outcomes

The use of expression data to predict disease states is quickly becoming standard practice. GeneSpring's k-nearest neighbors, and support vector machine technique allows the use of training sets to uncover patterns that discriminate between classes. GeneSpring not only provides methods to predict biological activity using patterns of gene expression, but it will also help you to identify genes that are optimal at making such predictions.

Characterize Novel Expression Patterns

GeneSpring gives you a broad choice of sophisticated methods for uncovering the most abundant patterns in your gene expression experiment, and understanding how these patterns are related, including:

- Hierarchical gene trees
- Hierarchical conditions
- Self-organizing maps
- k-means clustering
- QT clustering
- Principal Components Analysis (PCA)

GeneSpring leverages data found in publicly-available genomic databases to build gene ontologies based on the latest annotation information available. These ontologies provide insights about possible cellular activity associated with your expression data. GeneSpring also builds homology tables automatically, allowing you to compare expression patterns across multiple species or chip technologies.

Test Complex Hypotheses

Users can create complex experiments that link trends in expression data to a variety of test parameters, and can quickly refine hypotheses by re-computing previously defined analyses against different combinations of samples.

GeneSpring's data management algorithms permit even massive experiments to run on standard desktop hardware. Additionally, efficient data representation techniques permit dramatically faster reloading once an experiment is created. GeneSpring can be used to build experiments from a library of samples that represent:

- Target and reference expression profiles
- Responses to compound libraries
- Profiles from related studies
- Predicted clinical and experimental conditions

Ensure Data Quality

The GeneSpring global error model gives you a reliable estimate of expression measurement precision, especially when a limited number of replicate samples are available. This error model is independent of measurement technologies, adding power to GeneSpring's statistical tools. Additionally, 16 robust normalization setups are readily available, including:

- Intensity-dependent (LOWESS) normalization
- Global per-chip or per-gene normalization
- Normalization to specified control samples

Record and Automate Complex Analysis Tasks

GeneSpring allows you to automate the routine tasks of analyzing, interpreting, and archiving large volumes of expression data. By dragging and dropping icons to a canvas, you can record the steps of complex analyses in a simple script or combine existing scripts to create a more powerful script.

You can also share or exchange scripts with colleagues, maintain consistency in the analysis process, and simplify data management. Computationally-intensive analyses can easily be captured and off-loaded to Signet's central server farms via GeneSpring's remote execution feature.

Import and Compare Expression Data from Multiple Platforms

Using a simple drag-and-drop interface, you can import experiments created by any expression analysis technology – including Affymetrix, Amersham, Agilent, and catalog array data using a MIAME-compliant vocabulary.

Expression profiles of a given sample can be compared to all other samples, even samples that span multiple experiments, or are ones that are derived from disparate hardware technologies. GeneSpring guides users though the process of applying the normalization and scaling algorithms that are the most appropriate for the particular expression analysis platform.

Export Data and Images in Standard Formats

Images within GeneSpring can easily be exported in popular file formats that are readable by a number of applications. Virtually any graphical image can be exported in either PICT or PNG format to enhance the quality of articles for publication. GeneSpring can export expression data in MAGE-ML (Microarray Gene Expression – Markup Language) to facilitate exchange of data with other databases and applications.

Take Advantage of World-class Technical Support

Silicon Genetics' technical support team is available to respond promptly to your questions about expression analysis by phone or e-mail. Additionally, a broad assortment of training resources is available to speed up your learning process. Silicon Genetics offers a series of face-to-face workshops geared to accommodate beginning through advanced users. Online training sessions on a variety of topics are available – moderated in real-time by subject matter experts. For those preferring to learn on their own schedule, a collection of self-paced, interactive tutorials are accessible online.

GeneSpring Key Features

Advanced Statistical Tools

GeneSpring provides a host of tools that enable you to ask detailed questions about complex data sets. These include t-tests, one-way ANOVA tests, and post-hoc tests for reliably identifying differentially expressed genes. In addition, GeneSpring's class prediction tools can identify genes capable of discriminating between one or more experimental parameters or sample phenotypes. Groups of genes identified by expression profiling can be further characterized by performing searches for potential regulatory sequences.

Data Clustering

GeneSpring provides sophisticated clustering methods to uncover patterns of gene expression data and the relationships between these patterns. Researchers can use one technique, or a combination of clustering options, to characterize their data. These options include: gene trees (hierarchical clustering), experiment trees, self-organizing maps, k-means, Principal Components Analysis (PCA), and QT clustering.

Principal Components Analysis (PCA) allows you to reduce the complexity of your data by discovering a number of principal components that define most of the data variability. QT clustering is an unsupervised technique that allows you to specify both the minimum size and maximum correlation coefficient of each cluster in the analysis.

Visual Filtering

GeneSpring offers visually-intuitive filtering tools for both entry-level and advanced users. All visual filtering windows generate graphs of results in real-time. These filters allow researchers to exclude particular conditions, set minimum and maximum values, and choose specific gene lists to filter.

GeneSpring also has an advanced filtering window designed for power users. The advanced filtering window allows you to create complex Boolean expressions to identify genes with a highly-specific expression pattern. Once these filters are created, they can be saved and shared with other researchers via Signet.

3D Data Visualization

The 3D scatter plot tool provides in-depth and interactive representations of highly-complex data. Expression data values or analysis results can be placed on any of the three user-defined axes to create a powerful medium for array data presentation.

Data Normalization

Sixteen transformations are available for creating powerful and flexible normalization scenarios. Normalization steps can be applied in virtually any order, and include operations such as dye swapping experiments and median polishing. Scenarios can be saved and applied in other experiments.

Pathway Views

With the Pathway viewer, genes and their expression patterns can be visually characterized based on their location within a cellular pathway. Users can design their own pathway diagrams or directly import publicly-available pathway maps. Users can predict genes associated with discrete steps in the pathway of interest.

Search for Similar Samples

GeneSpring allows you to compare the expression profile of a given sample to all of the other samples in Signet or GeneSpring, even if they are derived from different experiments or aren't associated with any experiment at all.

Support for MIAME Compliance

MIAME is a standard that describes the minimal information that is needed to fully describe a gene expression experiment and being MIAME compliant is fast becoming a prerequisite for publishing your data. GeneSpring makes it easy to become MIAME-compliant. Customize MIAME-compliant attributes from an easy-to-use window in GeneSpring. Ensuring that your team adheres to MIAME guidelines is simpler than ever before.

Scripting

GeneSpring comes with comprehensive script building and editing capabilities. Researchers can create custom scripts to automate repetitive analytical tasks, ensure consistency in the analysis process and simplify data analysis management.

Using this tool, researchers can design scripts that automatically upload results to Signet, or combine scripts with basic functions to perform more complex analyses. GeneSpring also provides BioScript Library 2—a ready-made collection of scripts that span the complete analysis process. It includes scripts for automating a broad range of quality control, statistical analysis, and biological data query tasks.

PCA on Conditions

PCA on Conditions helps you identify groups of treatments that have a similar effect on the global expression profile by detecting the two or more principal components that govern the gene expression profiles. Principal components can be plotted on 2D and 3D scatter plots, allowing the scientist to detect clusters of treatments or assist in the quality control process by detecting outliers. The new condition scatter plot allows you to plot two

or three experimental parameter values for conditions (in 2D or 3D scatter plots), giving you a better insight into the distribution of the parameter values across your experiment.

Organism-Specific KEGG and GenMAPP Pathways Import

In addition to support for general pathways from the Kyoto Encyclopedia of Genes and Genomes, Silicon Genetics has added support for organism-specific pathways from KEGG. This allows you to view your expression data against the exact pathways in the organism under study. A number of GenMAPP pathways are also available, vastly expanding the number of pathways supported in GeneSpring.

What's New in This Release?

This section describes the new features in GeneSpring 7. Cross references to related chapters in this guide are also included.

Genome Import Wizard

The Genome Installation wizard in the prior release has been replaced with the Import New Genome wizard. Using the new wizard, you can:

- Download pre-made genomes from the Silicon Genetics mirror site
- Import genomes from files in the GenBank/EMBL DNA databases
- Create genomes from data in tab-delimited files

The Import New Genome wizard also provides the following improvements in genome and array import:

- Displays a summary of all contents of a genome
- Exports and imports genomes as zip files
- Provides a history function that records changes to genomes
- Supports sequence files in FASTA format
- Enables the selection of multiple GenBank gbk files to define a genome
- Accepts non-standard formats for annotations files or Master Table of Genes file
- Supplies a list of standard web links to add to the Gene Inspector window
- Enables previews of a genome before it is saved and then saves the genome to a particular location in GeneSpring's hierarchy
- Allows the inclusion of custom annotations, which can be used for web link searches
- Provides a checklist of things to do when saving a genome, including running GeneSpider, making gene lists from properties, building simplified ontologies, associating pathways with genomes, and building homology tables

Go to Chapter 4, "Importing Genomes" on page 4-1 for more information.

Project-based Navigation

To provide greater control over navigating, filtering, and managing data objects, GeneSpring 7 introduces the concept of a project. A project is a virtual container for data objects that represents the way data is structured in laboratory notebooks, internal reports, or proprietary databases. Projects can contain any data object, irrespective of the array technology (GeneSpring genome) with which the data objects are associated.

Projects enable you to tag data objects associated with an experiment using your own labels so they can be identified, searched, and filtered using the GeneSpring filtering functions. For example, projects let you:

- Determine what experiments were conducted to address specific research hypotheses
- Determine what samples were used in the creation of the experiments
- Identify data objects derived from the analyses of individual experiments

After data objects have assigned to a project, you can use filtering and search tools to locate the objects you want. The Show menu, located at the bottom of the main GeneSpring window, also lets you filter data by assigned projects or unassigned projects. Go to Chapter 6, "Managing Genomes" on page 6-1 for more information.

K-Nearest Neighbors Class Predictor

New to GeneSpring 7 is the Class Predictor algorithm, which replaces the Predict Parameter Values command. The new k-nearest neighbor classifier optimizes the settings for the training set of samples by applying a cross-validation algorithm to the number of nearest neighbors and the number of genes. The resulting output shows the prediction rate (number classified correctly), the number misclassified, and the number unclassified. Go to Chapter 15, "Clustering and Characterizing Data" on page 15-1 for more information.

Support Vector Machines

GeneSpring 7 introduces a new way to classify genes using gene expression data from DNA microarray experiments—Support Vector Machines (SVMs). An SVM is a learning algorithm that uses prior knowledge of gene function to identify unknown genes of similar function in expression data.

SVMs avoid several problems associated with unsupervised clustering methods, such as hierarchical clustering methods and self organizing maps. SVMs offer flexibility in choosing a similarity function, managing large data sets, handling large feature spaces, and identifying outliers. Go to Chapter 15, "Clustering and Characterizing Data" on page 15-1 for more information.

Master Table of Genes Editor

In GeneSpring 7, you can now modify genes and annotations in the Master Tables of Genes using the Master Table of Genes Editor. Go to Chapter 6, "Managing Genomes" on page 6-1 for more information.

MAGE-ML Export

Many journals have recently adopted a requirement that all authors describing gene expression data must submit the data into a public gene expression database. GeneSpring 7 now enables you to export your experiments in MAGE-ML, the Micro Array Gene Expression Markup Language.

When exporting MAGE-ML file formats for publication, you have the ability to map the gene identifiers in the MAGE-ML file to the composite sequence identifiers used by EBI. This feature enables you to export raw files, along with normalized data in MAGE-ML format. Sample attributes and parameters can also be exported. Go to Chapter 17, "Exporting Data Files" on page 17-1 for more information.

Basic Script Filters

Two new scripts have been added to the Basic Scripts function in GeneSpring 7. These scripts are:

- **Filter on Parameter**—Computes the correlation between an experimental parameter and expression-level data.
- **Filter on Sample Attribute**—Computes a correlation between a sample attribute value and expression-level data across all samples in an experiment.

Go to Chapter 16 , "Using Scripts, External Programs, and Plugins" for more information.

Find Significant Parameters

To identify genes that are correlated or uncorrelated to experimental parameters, GeneSpring 7 includes new tools for comparing expression data to a single numeric or non-numeric attribute or parameter. These tools enable you to find parameters that are correlated to an expression profile using analysis of variance, association tests, or simple correlation techniques. Go to Chapter 15, "Clustering and Characterizing Data" on page 15-1 for more information.

GeneSpring Java API

GeneSpring 7 contains a public JAVA that will allow users to add functionality to GeneSpring by developing small plugins using the JAVA programming language. The Application Programming Interface (API) enables the development of four different types of plugins:

- **External Program**—Specifies parameters for running programs interactively. The parameters include the name and location of the executable.
- **Database Preprocessor Plugin**—Specifies whether to preprocess data before importing it into GeneSpring.
- **Script Plugin**—Allows the development of completely new Script blocks.
- **Interactive Plugin**—Specifies parameters for running plugins interactively.

Go to Chapter 16 , "Using Scripts, External Programs, and Plugins" for more information.

Data Object Modification in Signet

GeneSpring 7 enables owners of an object to lock and modify data objects in Signet. Moreover, all members of a group who have access to the data object can also modify that object in Signet. Data objects that can be modified include experiments, interpretations, and samples. Go to Chapter 18, "Using Signet" on page 18-1 for more information.

2

GeneSpring Interface

This chapter introduces you to the GeneSpring interface. It includes the following topics:

- GeneSpring Window
- Menu Bar
- Navigator
- Genome Browser
- Colorbar
- Image Window
- Inspectors
- Wizards

GeneSpring Window

The GeneSpring window (Figure 2-1) provides access to an analytical workbench that enables you to visualize, organize, and manipulate genomic expression data. Experimental data from microarrays, Affymetrix chips, SAGE, or any technique that associates numbers with genes, can easily be imported into GeneSpring for rigorous analysis. Using the GeneSpring window, you can:

- Statistically analyze microarray data of both one-color and two-color experiments
- Graphically monitor analysis procedures and display results
- Update genome information from GenBank, UniGene, and LocusLink databases
- Annotate genes of interest based on the simplified gene ontology annotations
- Map genes to biochemical pathways and cytogenetic bands on chromosomes
- Generate gene homology tables across species

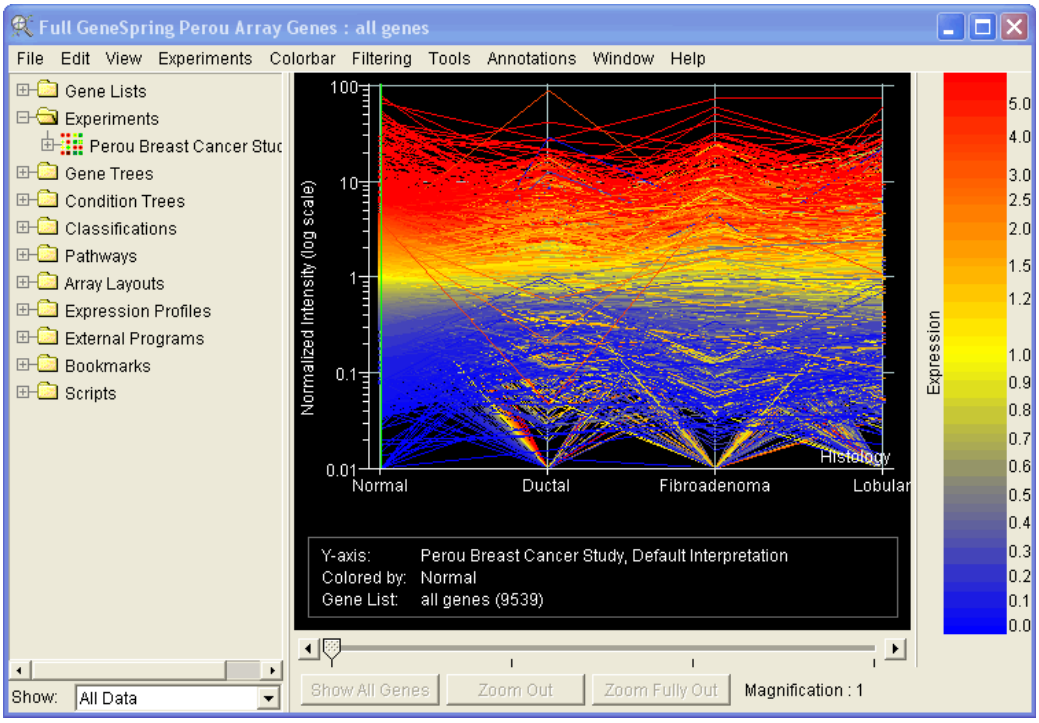


Figure 2-1 GeneSpring Window

Menu Bar

The menu bar, at the top of the GeneSpring window, provides access to menus and commands in GeneSpring.

File Menu

The File menu (Table 2-1) provides commands for managing data and images in GeneSpring.

Table 2-1 File Menu

Commands	Description
Import Data	Selects one data file for import.
Import Data from Database	Selects and imports data from a database.
Open Genome or Array	• Affymetrix —Opens an array from Affymetrix. • Demo Chip—Opens a genome from Silicon Genetics.
View Projects	• *<Project Name>*—Selects a project to view. • View Not Assigned—Shows objects, such as gene lists or experiments, that have not yet been assigned to a project.

Table 2-1 File Menu (Continued)

Commands	Description
Genome Manager	Opens, inspects, deletes, or renames genomes or arrays; logs in to Signet and uploads data; imports zip files into GeneSpring; and exports zip files from GeneSpring.
Import Genome	Imports a standard genome or creates a custom genome.
New Window	Creates a separate copy of the current window which allows you to work independently in that window. Changes you make in the new window are not implemented in the original window.
New Linked Window	Creates a linked copy of the current window which allows you to select one gene or gene list in two windows simultaneously. Changes you make in the linked window are reflected in the original window.
Log in to Signet	Logs in to Signet with a fully-qualified user name.
Bulk Upload to Signet	Uploads entire genomes or large amounts of data to Signet.
Copy Genome from Signet	Copies a genome from Signet and downloads it to GeneSpring.
New Pathway	• Import New Pathway—Imports an image file from your desktop or other server into GeneSpring. • Import KEGG Pathway—Imports a folder containing KEGG pathways from the KEGG FTP site into GeneSpring.
New Program	• New External Program—Specifies parameters for running programs interactively. • New Interactive Plugin—Specifies parameters for running plugins interactively. • New Script Plugin—Specifies parameters for running a script interactively. • New Database Preprocessor Plugin—Specifies whether you want to preprocess data before importing it into GeneSpring.
Import GeneSpring Zip	Imports a GeneSpring zip file.
Load Bookmark File	Loads a bookmark file created in GeneSpring.
Save Bookmark	Saves your current display settings, including experiment, gene list, coloration, and selected genes, in a bookmark.
Print Image	• Browser—Prints an image of the GeneSpring browser. • Browser and Colorbar—Prints an image of the GeneSpring browser and Colorbar. • Colorbar—Prints an image of the Colorbar panel only.
Save Image	• Browser—Selects the image file format, resolution, page size, label, and color scheme for an image you want to save. • Colorbar—Saves the Colorbar graphic as an image file.
Close	Closes the currently selected Genespring window.

Table 2-1 File Menu (Continued)

Commands	Description
Quit	Quits the current GeneSpring session and closes all open windows.

Edit Menu

The Edit menu (Table 2-2) provides commands for editing data in GeneSpring.

Table 2-2 Edit Menu

Commands	Description
Copy	• Copy Gene List—Copies the selected gene list. • Copy Experiment—Copies the selected experiment. • Copy Annotated Gene List—Copies the selected annotated gene list.
Paste	• Paste Gene List—Pastes the gene list. • Paste Classification—Pastes the selected classification (a grouping of genes by k-means or SOM clustering that is stored in the Classifications folder). • Paste Experiment—Pastes the experiment.
Undo	Reverses the last Edit command.
Edit Gene List	Filters on a gene list, creates a gene list, filters on an annotation, and shows all genes in a filtered list.
Find Gene	Finds a gene based on its systematic name, common name, or synonym.
Find Next	Finds the next gene that meets the defined search criteria.
Advanced Find Gene	Finds a gene using advanced search criteria.
Inspect Selected Gene	Displays detailed information about a gene, including its name, map, description, condition, and intensity. Also provides access to the Gene List Inspector window.
Search	Finds data by keyword, data by project assignment, data by advanced search criteria, or genes stored locally or on Signet.
Filter Navigator	Filters projects by keyword or other advanced search criteria.
Standard Attributes	Modifies the standard xml files for sample attributes and experimental parameters.
Preferences	Specifies global preferences for data files, database, colors, browsers, firewall, system, Signet use, computation, and other miscellaneous options.

View Menu

The View menu (Table 2-3) provides commands for displaying data in GeneSpring.

Table 2-3 View Menu

Commands	Description
Physical Position	Displays genes according to their physical position (when the gene loci are known and loaded into GeneSpring) within the DNA sequence of the organism.
Blocks	Displays a rectangle for every gene in the active genome, ordered by trust.
Tree	Displays the results of hierarchical clustering in the form of a mock phylogenetic tree or dendrogram. In such a tree, genes having similar expression patterns are clustered together.
Array Layout	Produces a synthetic picture of the arrays used in the current experiment. This view is useful in identifying arrays that display local shifts in intensity due to problems in probe deposition, hybridization, washing, or blocking.
Graph	Displays one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a line.
Graph by Genes	Displays an experiment as one line, where each point on the line represents the relative expression of one gene.
Bar Graph	Displays one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a vertical bar.
Pathway	Displays an imported gif or jpeg pathway image.
Ordered List	Displays a gene list in the order of its associated values.
Scatter Plot	Displays the expression levels of genes in two distinct conditions, samples, or normalization schemes.
3D Scatter Plot	Displays genes in a 3D scatter plot in which each dot represents a gene.
Compare Genes to Genes	Displays the similarity between the expression profiles of two genes in one list or in two separate lists.
Spreadsheet	Displays gene data in a spreadsheet.
Condition Scatter Plot	Displays the results of Principal Components Analysis performed on conditions.
Zoom In	Increases magnification step-by-step.
Zoom Out	Decreases magnification step-by-step.
Zoom Fully Out	Decreases magnification all at once.
Unsplit Window	Rejoins windows that have been split into several small windows.
Animate	Shows changes in classification assignments in real time.

Table 2-3 View Menu (Continued)

Commands	Description
Animate Secondary	Shows changes in classification assignments in real time when viewing two gene lists or experiments simultaneously.
Visible	• Picture—Shows or hides the optional picture at the bottom right corner of the main GeneSpring window. • Animation Controls—Shows or hides the slider and the Animate check box at the bottom of the window (hiding this check box does not disable the Animation feature). • Magnification—Shows or hides the Magnification feature and the Zoom Out button at the bottom of the window (hiding the Zoom Out button does not disable the Zoom Out menu option). • Secondary Picture—Shows or hides a secondary picture when viewing two gene lists or experiments simultaneously in the Genome Browser. • Secondary Animation Controls—Shows or hides the secondary Animation Controls check box and slider when viewing two gene lists or experiments simultaneously. • Navigator—Shows or hides the Navigator in the GeneSpring window. • Hide All—Hides everything in the GeneSpring window, except the Genome Browser. • Show All—Shows all elements. • Hide All in All Windows—Hides everything in all windows, except the Genome Browser. • Show All in All Windows—Shows all elements in all windows.
Remove Secondary Gene List	Removes a secondary gene list from the Navigator.
Display Options	Changes display options for views.
Show Average of Genes	Displays the centroid of all genes in the view.

Experiments Menu

The Experiments menu (Table 2-4) provides commands for managing experiment normalizations, parameters, interpretations, and samples.

Table 2-4 Experiments Menu

Commands	Description
Experiment Normalizations	Applies a transformation scenario to normalize data.
Experiment Parameters	Specifies the parameters to use in the experiment.
Cross-Gene Error Model	Applies a statistical model to estimate error in control and repeat measurements.
Experiment Interpretation	Specifies the type of analysis to apply, inspects experiment details, and displays the results. Also provides access to the Experiment Inspector.
Create New Experiment	Applies different filters to locate samples to include in a new experiment.
Duplicate Experiment	Copies an experiment and saves it under a new name.
Sample Manager	Applies different filters to locate samples.

Colorbar Menu

The Colorbar menu (Table 2-5) provides commands changing the current gene coloring scheme in the Genome Browser.

Table 2-5 Colorbar Menu

Commands	Description
Color by Expression	Colors genes according to their normalized expression values and trustworthiness.
Color by Significance	Colors genes according to the p-value from the t-test of differential expression.
Color by Venn Diagram	Colors genes based on their membership in one or more gene lists in a Venn diagram. • Make Gene Lists—Makes a Venn diagram based on a gene list. • Clear This Selection—Clears the gene list in the Venn diagram.
Color by Parameter	Colors genes based on the value of parameters. This scheme is best suited for use with Graph view and Bar Graph view when different conditions are indicated with discrete symbols.
Color by Classification	Colors genes based on a previously-defined classification scheme. You can use a folder of lists to color by classification or a classification method such as k-means or self-organizing maps.
No Color	Displays all genes as gray in color.

Filtering Menu

The Filtering menu (Table 2-6) provides commands for filtering data in GeneSpring.

Table 2-6 Filtering Menu

Commands	Description
Filter on Expression Level	Filters genes that have certain values present in some of the conditions or samples in an experiment or interpretation.
Filter on Fold Change	Filters genes based on a comparison of two samples or conditions.
Filter on Error	Filters data based on the standard deviation, standard error, or range of replicates.
Filter on Confidence	Filters data based on a measure of confidence, such as of confidence from the t-test p-value or number of replicates.
Filter on Parameters	Filters genes based on similar parameters.
Filter on Flags	Filters genes based on quality flags in the original data files.
Filter on Data File	Filters genes based on values in a specific column of the experiment data file.
Filter on Arbitrary File	Filters genes based on values in a specific column of any data file.
Filter on Gene List Numbers	Filters genes according to the numbers associated with them in a gene list, such as correlation coefficients, p-values, or fold change ratios.
Advanced Filtering	Provides advanced filters for more fine-grained filtering.

Tools Menu

The Tools menu (Table 2-7) provides commands for finding differentially expressed genes, finding genes with similar expression profiles, or clustering and characterizing genes.

Table 2-7 Tools Menu

Commands	Description
Statistical Analysis (ANOVA)	Compares mean expression levels between two or more groups of samples.
Clustering	• K-Means—Produces groups of genes with a high degree of similarity within each group, and a low degree of similarity between groups. • Gene Tree—Produced dendrograms that show the relationships among the different expression levels of genes over a series of conditions. • Condition Tree—Divides genes or conditions into groups that have expression patterns similar to nearby genes in a tree.

Table 2-7 Tools Menu (Continued)

Commands	Description
	• Self-Organizing Map—Divides genes into groups based on expression patterns and illustrates the relationship between groups by arranging them in a two-dimensional map. • QT Clustering—Identifies clusters of genes, such that each gene in the cluster is within a specified distance (based on a user-defined distance metric) of every other gene in the cluster.
Class Prediction	Predicts the value (or "class") of an individual parameter in an uncharacteristic sample or set of samples using k-nearest neighbors or Support Vector Machines techniques.
Find Potential Regulatory Sequences	Finds common regulatory sequences in a gene list, or searches for a known regulatory sequence.
Principal Component Analysis	Reduces a multi-dimensional data set by transforming a number of correlated variables into a smaller number of uncorrelated variables (principal components).
Find Similar Samples	Finds similar samples based on a gene list, target sample, or sample pool.
Find Significant Parameters	Finds parameters and attributes that are correlated to an expression.
Find Similar Genes	Finds similar genes based on a gene list and a similarity measure, such as a Pearson or Spearman correlation.
Draw Expression Profile	Draws a pseudo-gene to represent a hypothetical expression pattern.
Check Remote Execution Queue	Shows the status of all jobs: running, paused, pending, and completed, on Signet.
Script Editor	Creates re-usable scripts that can be applied to any data set.

Annotations Menu

The Annotations menu (Table 2-8) provides commands for adding, updating, and editing gene annotations.

Table 2-8 Annotations Menu

Commands	Description
GeneSpider	• Update Annotations from Silicon Genetics—Updates annotations from the Silicon Genetics database. • Update Annotations from GenBank—Updates annotations from the GenBank database. • Update Annotations from LocusLink—Updates annotations from the LocusLink database. • Update Annotations from UniGene—Updates annotations from the UniGene database.

Table 2-8 Annotations Menu (Continued)

Commands	Description
Build Homology Tables	Builds homologies between any pair of organisms that are included in both HomoloGene and UniGene, or within-genome homologies based on the UniGene Cluster ID or the LocusLink Locus ID.
Make Gene Lists from Annotations	Creates custom gene lists based on annotations.
Build Ontology	Hierarchically groups genes into meaningful biological categories (gene lists) based on the Gene Ontology Consortium Classifications (GO).
Edit Genes and Annotations	Displays the Master Table of Genes Editor which lets you modify genes and annotations. Also provides access to the Genome Inspector.

Window Menu

The Window menu (Table 2-9) provides command for managing windows in GeneSpring.

Table 2-9 Window Menu

Commands	Description
Full Window	Displays the full GeneSpring window.
Tile Windows	Arranges multiple GeneSpring windows in tile format on your desktop.

Help Menu

The Help menu (Table 2-10) provides access to supplementary information about in GeneSpring

Table 2-10 Help Menu

Commands	Description
Tutorial	Opens your default browser and takes you to the *GeneSpring Tutorial* in PDF format. You can save this file to your local machine and print it. The tutorial provides step-by-step procedures on how to use many basic feature and functions of GeneSpring.
User Manual	Opens the manual installed on your hard drive during installation or updating. The *GeneSpring User Manual* is a PDF document you can save or print.
Version Notes	Displays notes for your version of GeneSpring. The *versionnotes.html* are located in the `..\silicongenetics\genespring\docs` directory.
Keyboard Shortcuts	Opens a dialog box that shows the keyboard shortcuts for menus, tools, Views, 2D graphs, 3D graphics, and trees.
Update GeneSpring	Downloads the latest version of GeneSpring.

Table 2-10 Help Menu (Continued)

Commands	Description
Support and Training Resources	Displays the Silicon Genetics training page on the web site. Here you can take advantage of the many training options provided by our team.
Email Technical Support	Opens an email to request Silicon Genetics technical support.
Silicon Genetics Online	Opens the Silicon Genetics web site. This site contains a wealth of information including manuals and information on workshops designed to help you use GeneSpring more effectively.
Signet Database	Opens the Signet web browser. From here you can download a demo version of Signet and upload or download additional information. See the *Signet User Manual* for more information.
System Monitor	Displays the Java system monitor to track free memory and view the processes running on your computer.
Test Database Connectivity	Verifies that GeneSpring and the database are communicating with each other.
Show Helpful Hints	Displays a new helpful hint each time you start GeneSpring.
About	Displays information about GeneSpring, such as the version number and demo expiration date. If you contact Technical Support, they will ask you for this information.

Navigator

The Navigator (Figure 2-2) provides access to all data elements relating to genomes. Information is organized into files and folders. It appears in the left-hand panel of the GeneSpring main window.

Mini-navigators can also appear in many sub-windows in the GeneSpring interface. The files and folders that appear in the mini-navigators will differ depending on the sub-window that is displayed.

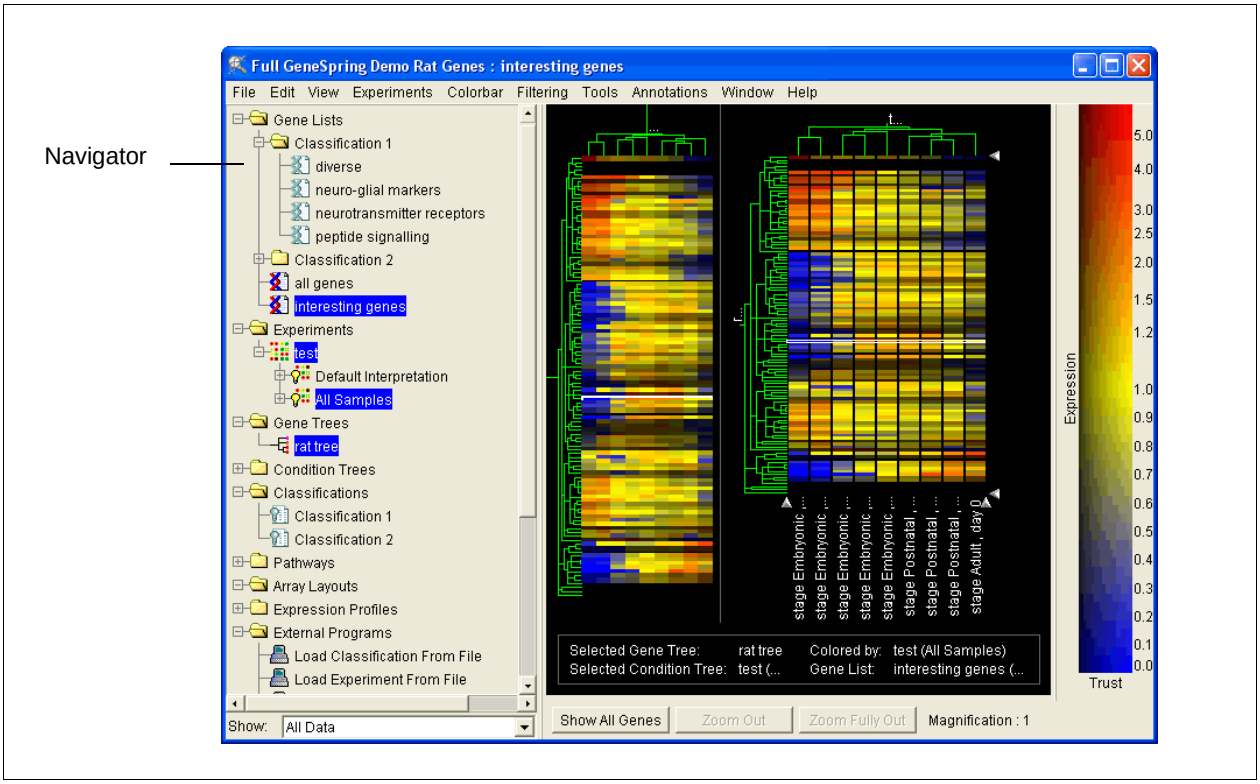


Figure 2-2 GeneSpring Window: Navigator

Folders

Navigator folders store information about different data objects in GeneSpring. Each folder contains a specific type of information. By default, folders in the Navigator are closed, although on start-up, GeneSpring displays an “all genes” or “all genomic elements” gene list. You can change the default genome that GeneSpring initially loads in the Preferences window, which is accessed through the Edit menu.

Gene Lists Folder

During analysis, you will create and work with interesting collections of genes known as gene lists. These gene lists are stored in the Gene Lists folder. By default, GeneSpring makes and displays an “all genes” list containing all genes in the genome.

Experiments Folder

The Experiments folder contains experiment information. Experiments are divided into *interpretations*. Experiment Interpretations tell GeneSpring how to treat and display your experiment variables, called experiment *parameters*.

Conditions are groupings of one or more samples. Each sample can be a condition, as in the “All Samples” interpretation, or a condition can include multiple samples. A condition can also include data from more than one sample.

Gene Trees Folder

Any gene trees created in GeneSpring are kept in the Gene Trees folder. Gene trees are dendrograms used as a method of showing relationships among the expression levels of genes over a series of conditions.

Condition Trees Folder

Condition trees are like gene trees, except that instead of showing the relationships between genes, they show the relationships among the expression levels of samples. Condition trees are kept in the Condition Trees folder.

Classifications Folder

The Classifications folder contains genes that have been grouped or *classified* into groups as defined by k-means or SOM clustering.

Pathways Folder

Pathways are images of regulatory or metabolic pathways that can be imported into GeneSpring. Genes are overlaid on these images allowing you to observe their changing expression levels across experimental conditions. A feature called, Find Genes Which Could Fit Here, can be used as a tool to predict new pathway elements.

Array Layouts Folder

The Array Layouts folder contains information about the arrangement of the spots on your array. These can be used to recreate an image of your arrays to check for regional abnormalities.

Expression Profiles Folder

Expression profiles are lines representing gene profiles that you draw in the Genome Browser. You can then search for genes matching that profile. Any expression profiles you create are stored in the Expression Profiles folder.

External Programs Folder

External programs are analysis programs outside GeneSpring that can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

Bookmarks Folder

Bookmarks are saved display settings such as an experiment, gene list, color scheme, or selected genes. You can always save your current display and return to it later by opening the Bookmarks folder and selecting a particular bookmark.

Scripts Folder

Scripts are tools that save time by allowing a long series of data analysis steps to be performed at once. Scripts are re-usable and can be applied to any data set. You can create your own scripts using the Silicon Genetics Script Editor. All scripts, including complimentary scripts shipped with GeneSpring, are stored in the Scripts folder.

Show Menu

The Show menu (Table 2-11), which appears at the bottom of the Navigator, provides quick access to commands for filtering data in the Genome Browser.

Table 2-11 Show Menu

Command	Description
All Data	Shows all data in the Genomic browser.
Not Assigned	Shows data that have not been assigned to a project.
Filter	Displays the Filter Navigator window which lets you search for data by keyword or other search criteria.

Menus

Right-clicking on a file or folder displays commands for performing different actions. Table 2-12 describes the commands that can appear. The commands can differ depending on the data object you select.

Table 2-12 Navigator: Menus

Command	Description
Display	Displays the contents of a file in the Genome Browser.
Attachments	Adds an attachment to the selected file.
Inspect	Opens an Inspector.
Upload to Signet	Uploads data from GeneSpring to Signet.
Copy from Signet	Copies data from Signet to GeneSpring.
Assign Project	Assigns a data object to a project.
Change Owner	Changes the owner of a project.
Rename	Renames a folder or file.
Delete	Deletes the selected folder or file.
Export as Zip	Exports the selected data as a zip file.
Clear (on folders only)	Clears the image in the Genome Browser.
Import Zip (on folders only)	Imports the selected zip file.
Add New Folder (on folders only)	Creates a new folder.

Special Menus

Right-clicking on certain files or folders displays special commands for performing different actions. Table 2-13 describes the commands that can appear.

Table 2-13 Navigator: Special Menus

Command	File/Folder	Description
Set as Default	Experiment	Automatically loads this experiment when GeneSpring starts.
Use as Coloring	Classification/Gene List folder	Colors genes based on some previously-defined knowledge about them.
Split Window	Classification/Gene List folder	Splits the Genome Browser into several small windows.
Display as Second Gene List	Gene List	Displays a second gene list in the Genome Browser for comparison.
Set Secondary Experiment	Experiment	Shows an additional experiment in the Navigator (only enabled for certain views).
Translate	Gene List/Experiment	Creates a homology gene list after logging in to Signet.
Re-Annotate	Gene Tree	Annotates genes using only standard lists.
Venn Diagram	Gene Lists	Creates a Venn diagram and colors a specific gene according to your specification.
Make Gene Lists	Classification/Pathways	Creates a list of genes.
Run	Script/External Programs	Runs a script or an external program.
Use as Coloring	Classification	Colors chromosome numbers by classification.
Use to Color Branches	Classification	Colors tree branches by classification.
Edit	Script	Edits a script.
Help	Script	Displays help for the selected script.
Duplicate	Experiment	Copies an experiment and saves it with a new name.
Save DB Data Locally	Experiment	Saves experiment information from a database to your local machine.
Export as MAGE-ML	Experiment	Exports data in MAGE-ML format.

Table 2-13 Navigator: Special Menus (Continued)

Command	File/Folder	Description
Draw Gene Profile	Expression Profile	Draws lines representing gene profiles in the Genome Browser.

Genome Browser

The Genome Browser (Figure 2-3), which appears in the center of the GeneSpring window, displays a visual representation of genes. You can zoom in on data using the mouse, pan around the data using arrow keys, or display the data in different views.

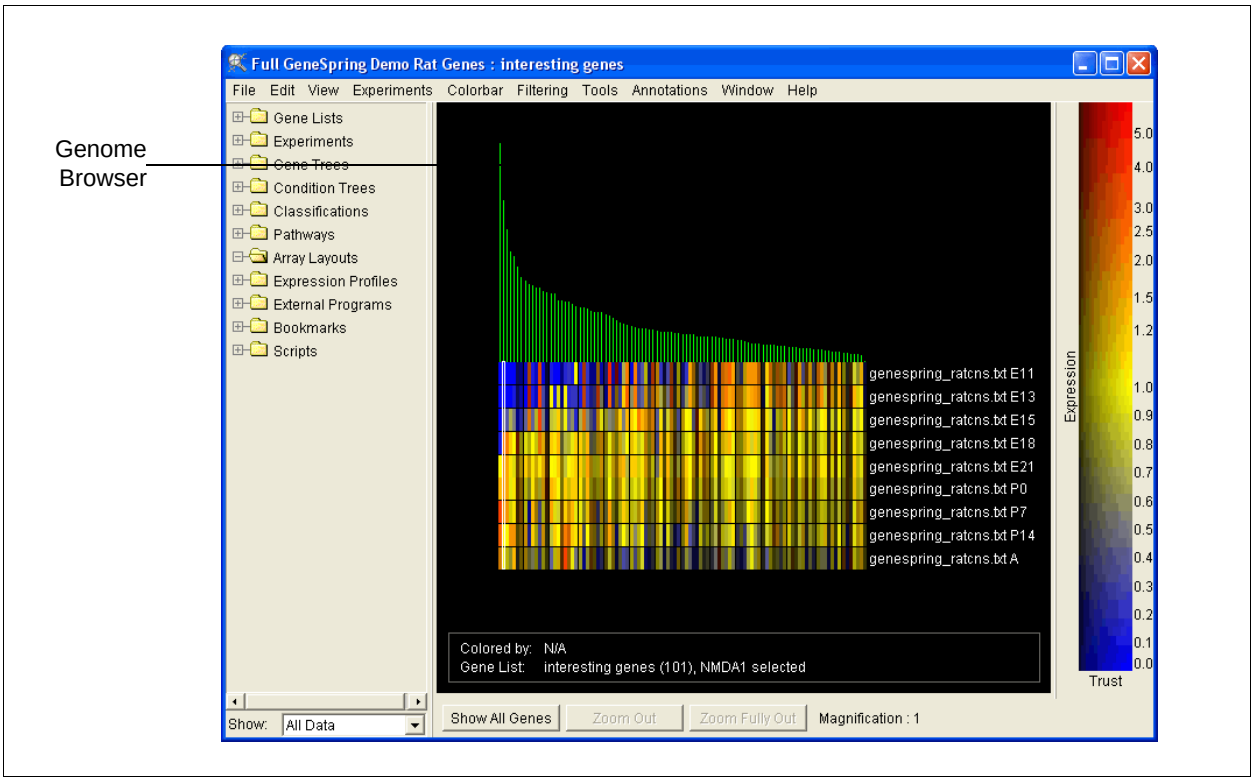


Figure 2-3 GeneSpring Window: Genome Browser

The View menu in the toolbar provides commands for displaying genomic expression data in different views. These views are:

- Physical Position
- Blocks
- Tree
- Array Layout
- Graph
- Graph by Genes
- Bar Graph

- Pathway
- Ordered List
- Scatter Plot
- 3D Scatter Plot
- Compare Genes to Genes
- Spreadsheet
- Condition Scatter Plot

Go to Chapter 8, "Viewing Data" for more information.

Menus

Table 2-1 describes the commands that are available when you right-click a data object in the Genome Browser.

Table 2-1 Genome Browser: Menus

Command	Description
Zoom In	Increases magnification step-by-step.
Zoom Out	Decreases magnification step-by-step.
Zoom Fully Out	Decreases magnification all at once.
Make List from Selected Genes	Makes a list of all the genes you have selected in the Genome Browser.
Save Expression Profile	Saves a new expression profile in the Expression Profiles folder in the Navigator.
Save Image	Saves an image file in pict or png format for export to other programs.
Save Bookmark	Save the current display settings, including experiment, gene list, color scheme, and selected genes, in a bookmark.
Display Options	Displays options for the selected view.

Special Menus

Table 2-2 describes pop-up commands that are available in certain views displayed in the Genome Browser.

Table 2-2 Genome Browser: Special Menus

Command	View	Description
Load Sequence	Physical Position	Loads nucleic acid sequences with genome data.
Make Subtree	Tree	Creates a new tree from a node of a larger tree.
Make List from Subtree	Tree	Makes a gene list from a subtree.
Display Subtree	Tree	Displays a subtree by clicking on a node.

Table 2-2 Genome Browser: Special Menus (Continued)

Command	View	Description
Display Parent of Subtree	Tree	Displays the tree immediately above the one selected.
Display Entire Tree	Tree	Displays the entire tree by clicking anywhere on the tree.
Find Genes Which Could Fit Here	Pathway	Predicts new pathway elements.
Delete Entire Tree	Tree	Deletes the selected tree.
Reset Orientation	3D Scatter Plot/Condition Scatter Plot	Resets the scatter plot to the default orientation.
Select Genes Within Lines	2D Scatter Plot	Selects and highlights genes within the fold lines.
Show Associated Values	Compare Genes to Genes	Shows the values from the selected gene list.
Hide Associated Values	Compare Genes to Genes	Hides the values from the selected gene list.
Unsplit Window	All views split by a classification	Rejoins windows that have been split into several smaller windows.
Remove Secondary Gene List	All views in which a secondary gene list is selected	Removes a secondary gene list from the Navigator.

Button Bar

The Button bar (Table 2-3), located at the bottom of the Genome Browser, provides functions for displaying data.

Table 2-3 Genome Browser: Button Bar

Button	Description
Show All Genes	Shows all genes in the Genome Browser.
Zoom In	Increases magnification step-by-step.
Zoom Out	Decreases magnification step-by-step.
Zoom Fully Out	Decreases magnification all at once.

Colorbar

The Colorbar (Figure 2-4) is located in the rectangle on the far right of the GeneSpring window. It is a two-dimensional representation of gene expression and trust. These measures are derived from the following genomic information:

- Signals that represent a gene or probe set in one-color experiments
- Signals that represent fluor intensities (usually Cy5 and Cy3) in two-color experiments
- Flags that represent gene intensity, trustworthiness of the measurements, or a combination of both.

Highly-reliable data is indicated by a high signal-strength value. Data of average reliability is indicated by a medium signal-strength value. Unreliable data is indicated by a low signal-strength value.

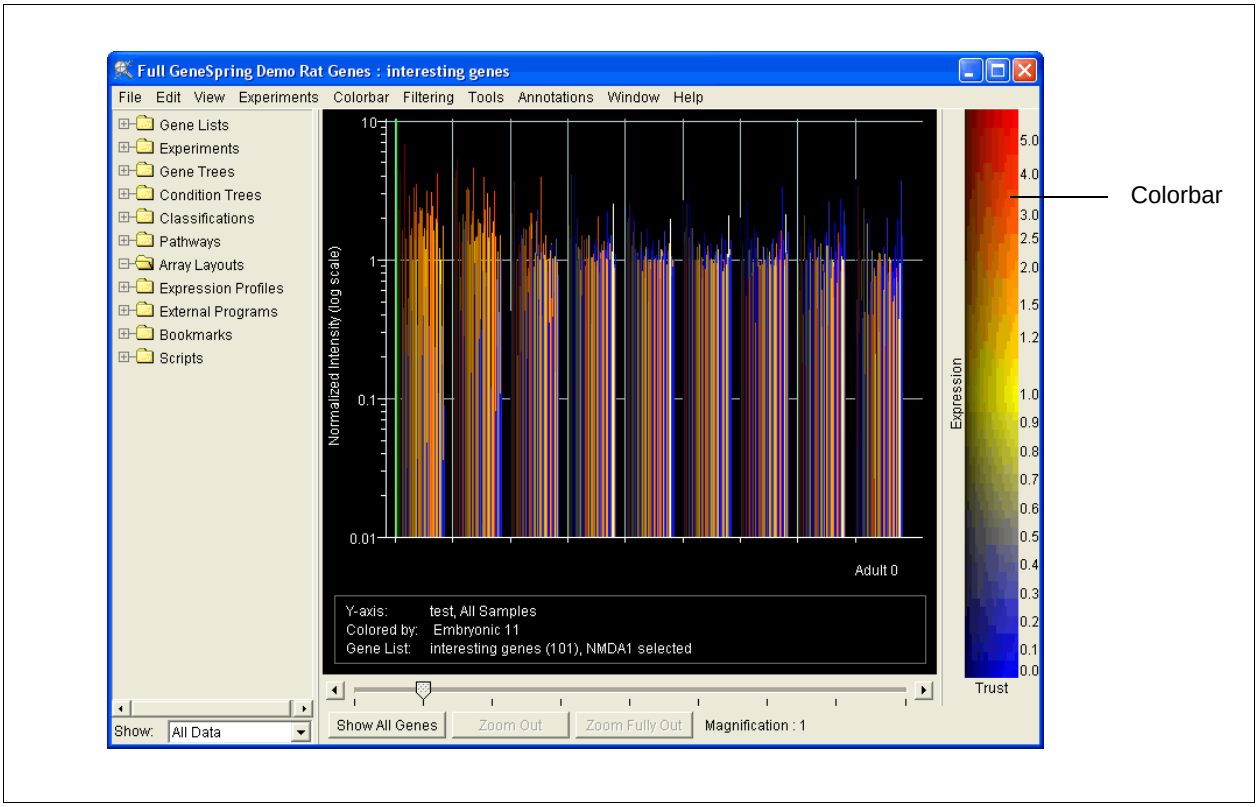


Figure 2-4 GeneSpring Window: Colorbar

Any gene with a signal strength above the value indicated as a high signal-strength will be colored using the brightest color appropriate. Any gene with a signal strength below the value given for unreliable data will appear black in color. The medium signal value gives the value for the mid-point of the Colorbar, and genes with a medium signal strength are colored halfway between the two color extremes.

The Colorbar menu in the toolbar also provides commands for coloring genomic data by expression, significance, Venn diagram, parameter, or classification. The Preferences

command in the File menu offers a choice of color schemes that can be used in the Colorbar.

Image Window

The Image window (Figure 2-5), located in the lower right corner of the GeneSpring window, displays images that correspond to the various points in an experiment. You can also drag the slider bar at the bottom of the Genome Browser to move to different points in your experiment. The corresponding data displays in the Image window.

The Animation command in the View menu also animates all points in an experiment, from start to finish, in the Image window. Many of these functions are also available in the View menu.

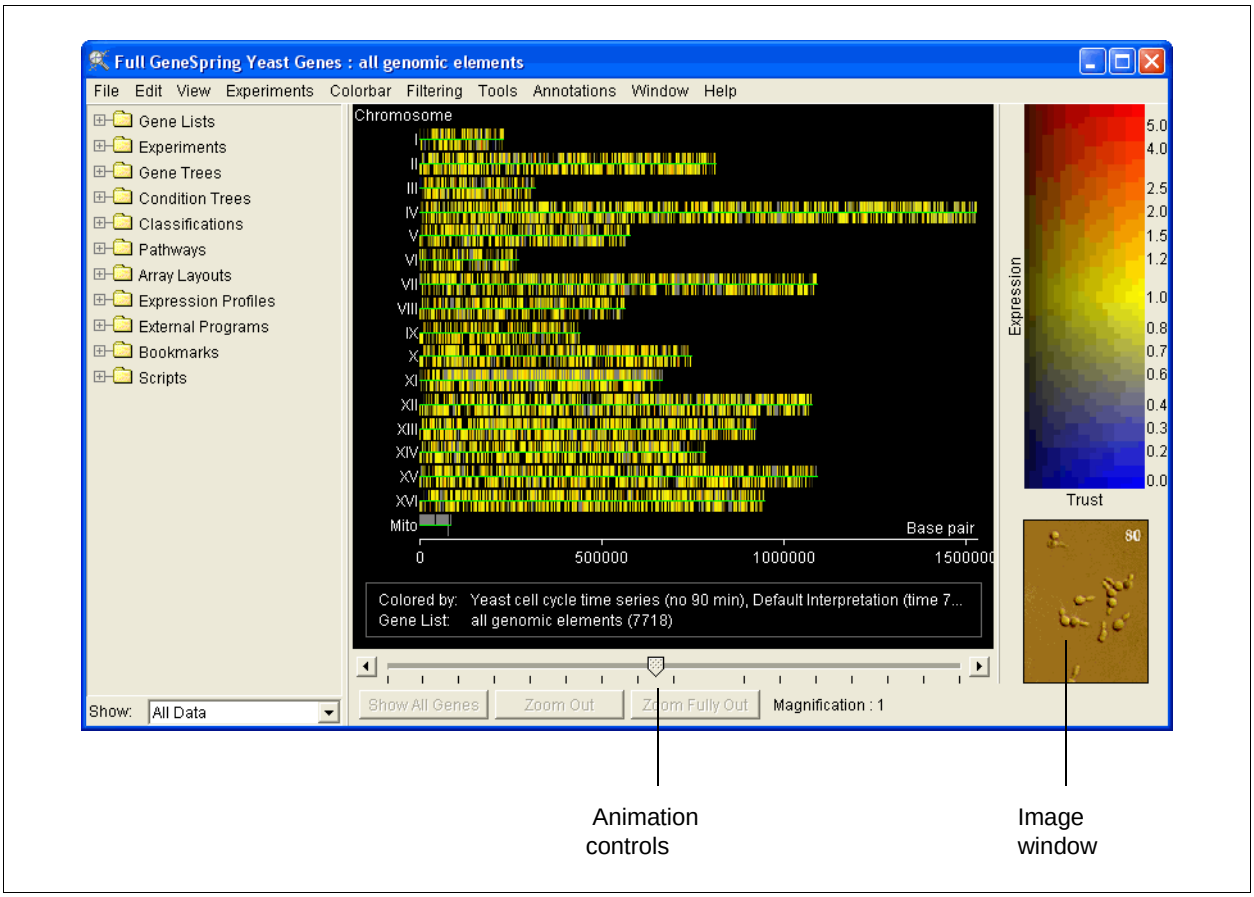


Figure 2-5 GeneSpring Window: Image Window and Animation Controls

Inspectors

Inspectors are special windows in GeneSpring that enable you to view the current defaults and available details of any gene, condition, classification, or experiment. The Gene Inspector (Figure 2-6) is one example of the many inspectors in GeneSpring.

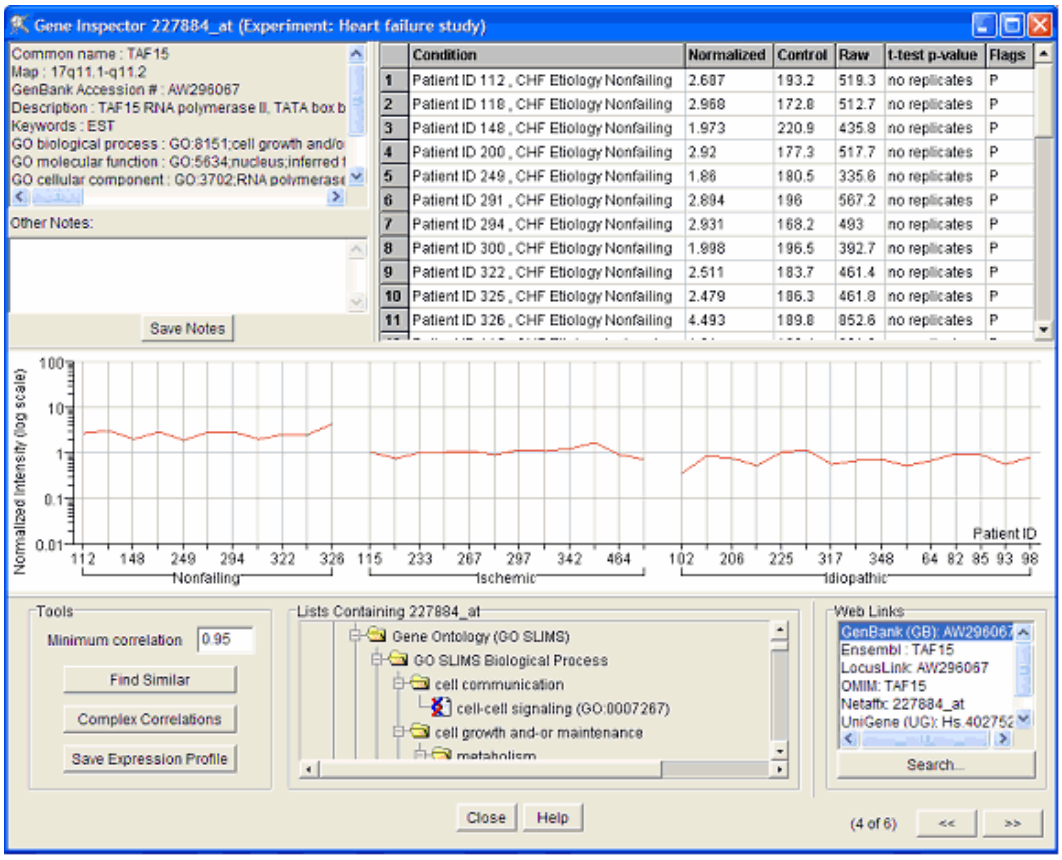

Gene Inspector 227884_at (Experiment: Heart failure study)
Common name : TAF15
Map : 17q11.1-q11.2
GenBank Accession # : AW296067
Description : TAF15 RNA polymerase II, TATA box b
Keywords : EST
GO biological process : GO:8151;cell growth and/o
GO molecular function : GO:5634;nucleus;inferred f
GO cellular component : GO:3702;RNA polymerase
Other Notes:
Save Notes
Condition
Normalized
Control
Raw
t-test p-value
Flags
1 Patient ID 112 , CHF Etiology Nonfailing 2.687 193.2 519.3 no replicates P
2 Patient ID 118 , CHF Etiology Nonfailing 2.968 172.8 512.7 no replicates P
3 Patient ID 148 , CHF Etiology Nonfailing 1.973 220.9 435.8 no replicates P
4 Patient ID 200 , CHF Etiology Nonfailing 2.92 177.3 517.7 no replicates P
5 Patient ID 249 , CHF Etiology Nonfailing 1.86 180.5 335.6 no replicates P
6 Patient ID 291 , CHF Etiology Nonfailing 2.894 196 567.2 no replicates P
7 Patient ID 294 , CHF Etiology Nonfailing 2.931 168.2 493 no replicates P
8 Patient ID 300 , CHF Etiology Nonfailing 1.998 196.5 392.7 no replicates P
9 Patient ID 322 , CHF Etiology Nonfailing 2.511 183.7 461.4 no replicates P
10 Patient ID 325 , CHF Etiology Nonfailing 2.479 186.3 461.8 no replicates P
11 Patient ID 326 , CHF Etiology Nonfailing 4.493 189.8 852.6 no replicates P
Normalized Intensity (log scale)
100
10
1
0.1
0.01
112 148 249 294 322 326 115 233 267 297 342 464 102 206 225 317 348 64 82 85 93 98
Patient ID
Nonfailing
Ischemic
Idiopathic
Tools
Minimum correlation 0.95
Find Similar
Complex Correlations
Save Expression Profile
Lists Containing 227884_at
Gene Ontology (GO SLIMS)
GO SLIMS Biological Process
cell communication
cell-cell signaling (GO:0007267)
cell growth and-or maintenance
metabolism
Web Links
GenBank (GB): AW296067
Ensembl : TAF15
LocusLink: AW296067
OMIM: TAF15
Netaffx: 227884_at
UniGene (UG): Hs.402752
Search...
Close
Help
(4 of 6)
<<
>>

Figure 2-6 Gene Inspector Window

Table 2-4 describes the different inspectors that are available in GeneSpring. Refer to Chapter 7, "Finding, Selecting, and Inspecting Data" for detailed information.

Table 2-4 GeneSpring Inspectors

Inspector	Description
Classification	Displays the methods used to construct a classification.
Condition	Displays all conditions associated with a sample.
Experiment	Displays all data associated with an experiment.
External Program	Displays the parameters specified for running programs interactively.
Gene	Displays all data associated with a gene.
Gene List	Displays the contents of a gene list.

Table 2-4 GeneSpring Inspectors (Continued)

Inspector	Description
Genome	Displays all the data associated with a genome.
Inspect Selected Gene	Displays the Gene Inspector for the selected gene.
Interpretation	Displays all interpretations associated with an experiment.
Sample	Displays all data associated with a sample.
Script	Displays properties associated with scripts, building blocks, history, and change information.

Wizards

Wizards are special programs in GeneSpring that guide you, step-by-step, through more complex procedures. One such wizard is the Import Genome wizard (Figure 2-7) which guides you through the process of downloading a standard genome or creating a custom one.

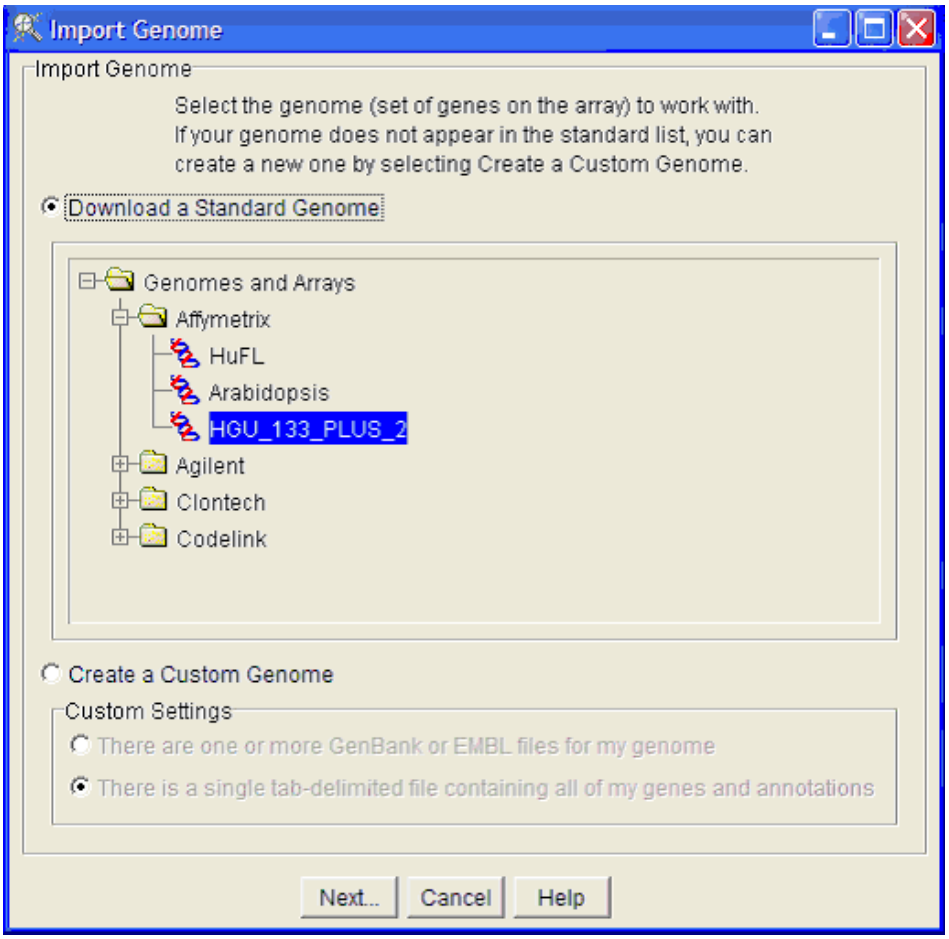

Import Genome
Import Genome
Select the genome (set of genes on the array) to work with.
If your genome does not appear in the standard list, you can
create a new one by selecting Create a Custom Genome.
Download a Standard Genome
Genomes and Arrays
Affymetrix
HuFL
Arabidopsis
HGU_133_PLUS_2
Agilent
Clontech
Codelink
Create a Custom Genome
Custom Settings
There are one or more GenBank or EMBL files for my genome
There is a single tab-delimited file containing all of my genes and annotations
Next...
Cancel
Help

Figure 2-7 Import Genome Wizard

Table 2-5 describes the wizards that are available in GeneSpring.

Table 2-5 GeneSpring Wizards

Wizard	Description
Import Data	Imports a tab-delimited text file from a local system or network server.
Import Data from Database	Imports data from an external database.
Import GeneSpring Zip	Imports a GeneSpring zip file from a local system or a network server and uncompresses the contents.
Import Genome	Downloads a standard genome or creates a custom one.
Experiment	Creates a new experiment and specifies experiment parameters, normalizations, and interpretations.
Pathway	Imports a new pathway from an external FTP site.

3

Getting Started

This chapter explains how to get started with GeneSpring. It covers the following topics:

- Requirements
- Installing GeneSpring
- Starting GeneSpring
- Working with Previous Releases of GeneSpring
- Updating GeneSpring
- Learning to use GeneSpring
- GeneSpring Basics
- Setting Preferences

Requirements

This section lists the requirements for the GeneSpring supported platforms.

Windows

- Windows 2000/XP
- Pentium II or better
- 256-MB RAM (512-MB recommended)
- 1024 x 768 display
- 40-MB disk space

Macintosh

- MacOSX.2 or higher
- Power PC or better
- MRJ 2.2.5
- 256-MB RAM (512-MB recommended)
- 1024 x 768 display
- 40-MB disk space

UNIX

- Solaris 8/9 or Redhat Linux Enterprise 3.0
- JDK1.4.2_0.4 or later
- 256-MB RAM (512-MB recommended)
- 1024 x 768 display
- 40-MB disk space

Installing GeneSpring

Two methods are available for installing GeneSpring. You can install the application from CD or download it from the Silicon Genetics web site.

Installing GeneSpring from CD

To install GeneSpring from CD:

1. Do one of the following:
   - Select **Install GeneSpring**.

     A splash window and an InstallAnywhere window opens with a progress bar.
   - (Windows only) Select **Start** > **Run** in the Start menu. Enter `D:\gspring.exe`, where D is the CD-ROM drive on your computer.
2. Follow the on-screen instructions. For more information see the *ReadMe* file included with the CD.

Installing from the Web

If you are reading this manual and do not have a copy of GeneSpring, you can download a copy by going to the following URL:

http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/download.smf

Follow the on-screen instructions and Silicon Genetics will send you a user name, password, and download link.

Starting GeneSpring

To start GeneSpring:

- Double-click the GeneSpring icon.

- (Windows) Select **Start > Programs > GeneSpring** and double-click the GeneSpring program.
- (Macintosh) Go to the Applications/Silicon Genetics directory and double-click the GeneSpring program.

A splash window opens containing your GeneSpring version number, the expected expiration date, and the JVM you are using. You will then see the GeneSpring main window. For further details, see "GeneSpring Basics" on page 3-5.

Obtaining a License Key

If you have already installed a demo copy of GeneSpring, your license key will expire within one month of the initial installation. Once you have purchased a full GeneSpring license, Silicon Genetics will send you a license key. Save this license key file in the `..\silicongenetics\genespring\data` folder.

If you have kept the default settings of GeneSpring, go to the Program Files directory on Window or the Applications folder on Mac. When the key is about to expire, you will get a warning message 30 days in advance.

If your license has expired, or it is about to expire, contact the Silicon Genetics at techsupport@silicongenetics.com or call us at +1-(866) SIG SOFT(866) 744 7638 (in the US) or +81 (70) 5072-2085 (Japan) or +44 (0) 1259 751 833 (Europe). Support via email is also available at: support@silicongenetics.com.

Setting Memory Usage Options

Once GeneSpring is installed, ensure that the default memory setting in GeneSpring preferences is set to approximately half of your computer's available memory (or more if you have a lot of RAM).

To set memory usage:

1. Select **Edit > Preferences**, and then choose **System.**
2. In the **Desired Memory Use** field, enter the amount of memory you want to use.

Verifying Virtual Memory

At least 150 MB of virtual memory is required for optimal GeneSpring performance. To ensure that large files are not interfering with software performance, you may need to move some large files to a different hard disk.If you continue to experience slow performance, you need to check your memory usage.

To verify virtual memory:

1. Select **Help > System Monitor** before invoking any functions.
2. Make a record of the Total Memory and Free Memory listed in the System Monitor window.
3. Contact Silicon Genetics at techsupport@silicongenetics.com or call us at +1-(866) SIG SOFT(866) 744 7638 (in the US) or +81 (70) 5072-2085 (Japan) or +44 (0) 1259 751 833 (Europe). Support via email is also available at: support@silicongenetics.com.

Working with Previous Releases of GeneSpring

GeneSpring 7 supports the conversion of data formats created before release 5.0. When you open data formats created in a previous release of GeneSpring, the Old Data Format window opens. You will be asked to confirm or abort the conversion to the latest release.

Updating GeneSpring

To update an existing GeneSpring installation:

1. Select **Help > Update GeneSpring**.
2. Follow the on-window instructions to obtain the current GeneSpring version.

Learning to use GeneSpring

This section describes the resources that are available to you for learning GeneSpring. The Help Menu, located on the GeneSpring window menu bar, provides access to resources for learning to use GeneSpring.

Tutorial

The Tutorial command opens your default browser and takes you to the GeneSpring Basics Instructional Manual in PDF format. You can save this file to your local system and print it. The tutorial covers many basic topics of GeneSpring.

User Manual

Select the User Manual command to open the manual installed on your hard drive during installation or updating. The *GeneSpring User Manual* is a PDF document you can save or print.

Version Notes

Select the Version Notes command to view notes for your version of GeneSpring. These are located in `..\silicongenetics\genespring\docs\` directory.

Update GeneSpring

Select the Update GeneSpring command to download the latest version of GeneSpring. You must have an active license key to update your software.

Technical Support

Select the Technical Support command to contact the Silicon Genetics technical support team by email.

Silicon Genetics on the Web

Select the Silicon Genetics on the web command to browse the Silicon Genetics web site. This site contains a variety of information including manuals and information on workshops designed to help you use GeneSpring more effectively.

Signet Database

The Signet Database command redirects you to a web page that provides information about Signet.

Support and Training Resources

Select the Support and Training Resources command to view the Silicon Genetics training page. Here, you can take advantage of the many training options provided by Silicon Genetics.

System Monitor

Select the System Monitor command to view the Java system monitor to track free memory and view the processes running on your computer.

Test Database Connectivity

If you are using a database with GeneSpring, select the Test Database Connectivity command to verify that GeneSpring and the database are communicating with each other.

Show Helpful Hints

Select the Show Helpful Hints command to display a new helpful hint each time you start GeneSpring.

About

Select the About command to view information about GeneSpring such as the version number and demo expiration date. If you contact technical support, you will be asked to supply the version number of GeneSpring you are using.

GeneSpring Basics

This section provides a brief introduction to GeneSpring basics. It is designed to familiarize you with the workflow, terminology, actions, and functions of GeneSpring.

Workflow

Figure 3-1 depicts steps that can occur in a typical data analysis session using GeneSpring. Although these steps do not represent all the functions that can be performed in GeneSpring, use the information as a guideline for your workflow.

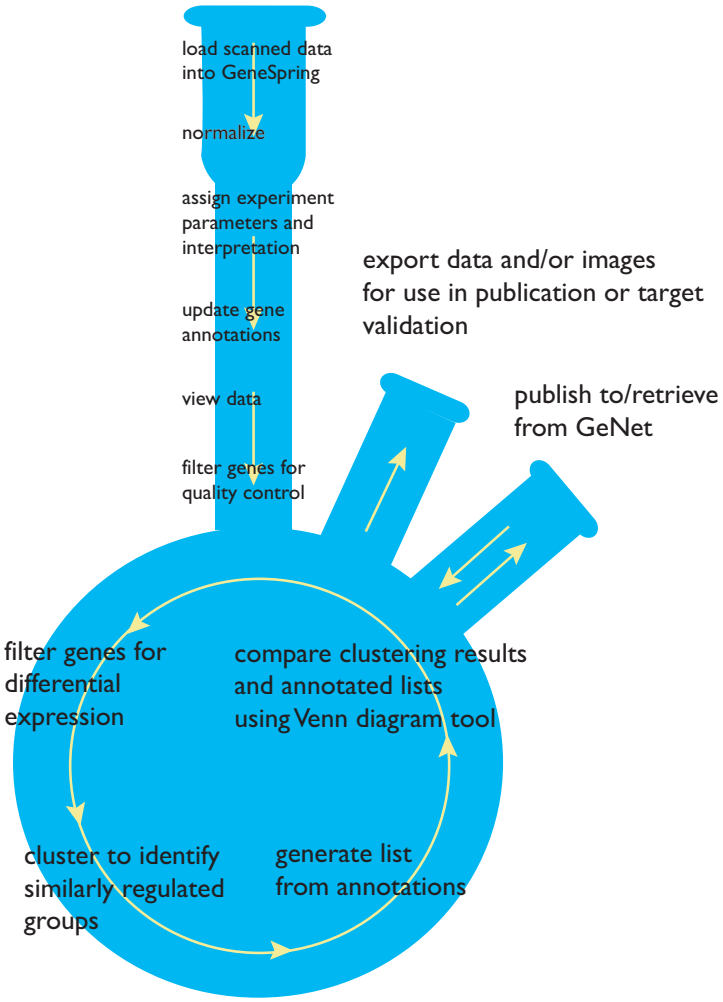

load scanned data into GeneSpring
normalize
assign experiment parameters and interpretation
update gene annotations
view data
filter genes for quality control
export data and/or images for use in publication or target validation
publish to/retrieve from GeNet
filter genes for differential expression
compare clustering results and annotated lists using Venn diagram tool
cluster to identify similarly regulated groups
generate list from annotations

Figure 3-1 GeneSpring Workflow

Terminology

In the process of working with data in GeneSpring, you will encounter new terminology. Below are explanations of how these terms are used in GeneSpring.

What is a Genome?

In the context of GeneSpring, a genome contains information about all the genes in your chip or microarray setup. Note that a GeneSpring genome does not correspond exactly to the biological definition of a genome. A genome in GeneSpring is composed of discrete genes, as opposed to the full nucleotide sequence. This means that a GeneSpring genome can contain two genes representing alternately spliced variants of a single gene, whereas a true genome would include the DNA sequences for only one.

What is a Parameter?

Parameters are experiment variables, such as age, weight, LDL level, blood pH, or IQ. Parameter values are values assigned to experiment parameters. For example, Embryonic, Postnatal. or Adult could represent parameter values of the experiment parameter *stage*, while .01 ppm could represent a parameter value of the experiment parameter *concentration*.

What are Replicates?

Replicates are repeated experiments with the same sample. They provide a measure of the experimental variation, such as:

- multiple spots on the same array representing the same gene (also referred to as a copy)
- the same sample on more than one array
- a biological replicate (equivalent samples taken from more than one organism)

Graphically, a parameter defined as a replicate is a hidden variable; no visual distinction is made based on this parameter or its parameter values.

What is a Region?

Regions divide your data into specific sections. This is important if you use multiple arrays and want to normalize sections of an array separately, rather than normalizing across the entire data set.

What is Raw Data?

The analysis process begins by obtaining data in the form of flat files that were generated by your scanning software or other expression analysis technology. GeneSpring is capable of recognizing most commercially-available formats and can be customized to work with other formats, as necessary. Typically, the gene, spot, or probe-set intensity values in these files are referred to as raw data.

What is Normalized Data?

If GeneSpring recognizes your file format, it applies a set of default normalizations appropriate for your expression analysis technology. The denominator used to normalize each measurement is referred to as the control strength.

What is Interpreted Data?

GeneSpring can interpret normalized data in many different ways. You can elect to have multiple samples treated as replicates and averaged, and indicate what assumptions you GeneSpring should make about the precision of these averaged values. You can display and perform analyses on normalized data using three modes: ratio, log of ratio, or fold change.

It is important to note that the graphical display of normalized values and the numbers used for all analyses (such as clustering) reflect the mode you have chosen. However, the numbers displayed as text (as in the Gene Inspector window) and entered by the user as parameters for analyses (as in the Filter Genes tools) are always in ratio mode.

What are Flags?

Flags are additional measurement markers in your data set that can be used in subsequent analyses. They can be assigned as present, marginal, unknown, or absent.

Basic Actions

Once you have loaded your data, GeneSpring opens a window containing information from your new genome (Figure 3-2).

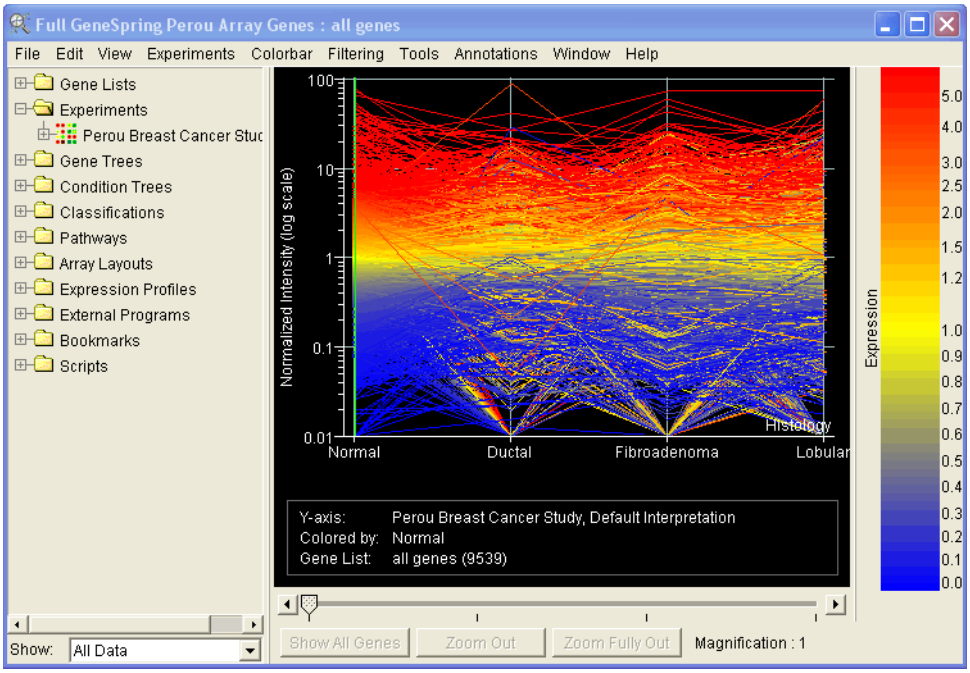


Figure 3-2 GeneSpring Window

Initially, all the genes in your experiment are displayed. To see your new genome select **File > Open Genome or Array** and choose your genome from the pop-up menu.

The following section describes some basic procedures for navigating the GeneSpring interface.

Changing the Genes Displayed

Open the gene list folder in the Navigator. GeneSpring initially displays the “all genes” list. You can change the genes shown in the display by choosing another list.

Changing Views

You can change the view in the Genome Browser using the View menu. GeneSpring initially displays the Classification view, in which genes are displayed according to pre-defined categories. However, you can also view displayed genes as a graph, a scatter plot, a bar graph, an ordered list, etc.

Note that some views such as Tree, Pathway, and Array Layout require some preparation, such as creating a tree or adding a pathway or array layout image. For details on views, see Chapter 8,"Viewing Data".

Zooming In

To zoom in on a region or gene, click on an area and drag your cursor diagonally. An expanding rectangle appears. Release the mouse and GeneSpring zooms in on the region enclosed by this rectangle.

Zooming Out

To zoom out, click **Zoom Out** or right-click (Ctrl + click for Mac) and choose **Zoom Out** to go back one level or **Zoom Fully Out** to zoom out as far as possible.

Moving Around the Window

You can move around a magnified view of a window by using the Page Up, the Page Down, or the arrows keys.

Selecting a Gene

Click once on a single gene to select it.

Selecting Multiple Genes

Hold down the Shift key and drag the mouse to select multiple genes. Alternatively, hold down the Shift key and click on individual genes to select them one by one.

Finding a Specific Gene

Select **Edit > Find Gene** or **Ctrl+F**. Enter the gene name or keyword and click the OK button. GeneSpring selects and zooms in on the gene.

Inspecting Genes

You can view detailed information about a gene by double-clicking on it to bring up the Gene Inspector window. This is easier after zooming in on the gene. A shortcut to the Gene Inspector is Ctrl + I, or ⌘+I for Macintosh users.

Reversing an Edit Command

You can undo your last action by selecting **Edit > Undo** or **Ctrl + Z** (⌘ + Z for Macintosh users).

Making your First Gene Lists

To make your first gene lists:

1. Select **Annotations > Make Gene Lists from Annotations**.
2. Choose the property you want to use for generating lists.
3. Click **OK**.

To make a list based on biological function:

1. Select **Annotations > Build Ontology**.
2. Name your new list.
3. Click **OK**.

To make lists from a group of selected genes:

1. Right-click over a highlighted group of genes.
2. Select **Make List from Selected Genes** from the menu.

Your new lists appear in the Gene Lists folder.

Drag-and-Drop Actions

The following drag-and-drop actions are supported in GeneSpring:

- Drag and drop external files into GeneSpring
- Drag and drop files out of GeneSpring
- Drag and drop into the Venn Diagram
- Drag and drop within Navigator
- Drag and drop gene lists into GeneSpring
- Drag and drop data files to invoke the Autoloader function

Mouse and Keyboard Actions

Table 3-1 describes the mouse and keyboard actions that are available in GeneSpring.

Table 3-1 Mouse and Keyboard Actions

To . . .	Do this . . .
Select a gene	Move the pointer on the gene and click the mouse button once.
Select non-adjacent genes	Move the pointer on the gene and click the mouse button once. Then, move the pointer over the next gene, hold down Shift, and single-click the mouse.
Display the Gene Inspector	Move the pointer on the gene and double-click the mouse.
Zoom In	Increases magnification step-by-step.
View a sub-tree (Tree view)	Double-click the node in the tree.
View nearby nodes in a left-hand tree (Tree view)	Hold down Ctrl+ an arrow key.
View nearby nodes at the top of the tree (Tree view)	Hold Alt+ an arrow key.
Select a single node in a tree (Tree view)	Single-click the node in the tree.

Table 3-1 Mouse and Keyboard Actions (Continued)

To . . .	Do this . . .
Move the 3D scatter plot axes (3D Scatter Plot view)	• To rotate the three-dimensional image—Hold down the Ctrl (Alt on Macintosh) key and then click and drag the mouse. • To rotate around specific axes—Hold down the X, Y, or Z keys. • To rotate in the opposite direction—Hold down the X, Y, or Z keys and the Alt key at the same time. • To rotate faster—Hold down the X, Y, or Z keys and the Shift key at the same time.
Change the 2D scatter plot axes (Scatter Plot view)	Right-click the scatter plot, select Display Options, and then select the Horizontal Axis or Vertical Axis tab.

Keyboard Shortcuts

Table 3-2 describes the keyboard shortcuts that are available in GeneSpring.

Table 3-2 Keyboard Shortcuts

To . . .	Do this . . .
Menus and Tools	
Import data	Ctrl+O
Undo an Edit command	Ctrl+Z
Find a gene	Ctrl+F
Find the next gene	Ctrl+G
Perform an advanced gene find	Ctrl+Shift+F
Inspect the selected gene	Ctrl+I
Save a browser image	Ctrl+B
Close a window	Ctrl+W
Quit GeneSpring	Ctrl+Q
Views	
Physical Position	Ctrl+‘
Blocks	Ctrl+1
Tree	Ctrl+2
Array Layout	Ctrl+3
Graph	Ctrl+4
Graph by Genes	Ctrl+5
Bar Graph	Ctrl+6
Pathway	Ctrl+7
Ordered List	Ctrl+8
Scatter Plot	Ctrl+9
Compare Genes to Genes	Ctrl+0

Table 3-2 Keyboard Shortcuts (Continued)

To . . .	Do this . . .
3D Scatter Plot	Ctrl+-
2D Graphics	
Zoom In	Ctrl+]
Zoom Out	Ctrl+[
Zoom Fully Out	Ctrl+HOME
Pan Around when Zoomed In	Arrow keys
Scroll Up when Zoomed In	Page Up
Scroll Down when Zoomed In	Page Down
Draw/Hide Expression Profile	Ctrl+D
Change Shape of Expression Profile	Ctrl+click
3D Graphics	
Rotate	Ctrl+drag
Rotate Clockwise	X,Y,Z
Rotate Counter-clockwise	Alt+X,Y,Z
Rotate Faster Clockwise	Shift+X,Y,Z
Rotate Faster Counter-clockwise	Alt+Shift+X,Y,Z
Left Hand Tree	
Parent Node	Ctrl+Left Arrow
First Child Node	Ctrl+Right Arrow
Sibling Node Below	Ctrl+Up Arrow
Sibling Node Above	Ctrl+Down Arrow
Right Hand Tree	
Next Left Sibling Node	Alt+Left Arrow
Next Right Sibling Node	Alt+Right Arrow
Parent Node	Alt+Up Arrow
First Child Node	Alt+Down Arrow

Tips for Macintosh Users

Except where otherwise noted, instructions in this manual describe GeneSpring usage on a PC. If you are a Macintosh user, you may find the following keystroke and mouse conversion information helpful:

- **Right-Click—**Hold Ctrl and click. This action most often activates a menu.
- **Ctrl = ⌘—**Substitute the ⌘ key wherever the manual mentions Ctrl. For example, if the manual says "press Ctrl + I to reach the Gene Inspector," substitute the ⌘ (Apple) key for Ctrl.
- **Drawing genes on a pathway—**Hold down the Option key and drag your cursor diagonally to draw a gene on a pathway. See "Changing Common Display Options" on page 8-57 for more information.

Note that on a Macintosh the menu bar is at the top of the window, not on the individual GeneSpring windows as displayed in this manual.

Basic Functions

This section describes commonly-used functions in GeneSpring.

Opening a Different Genome

To open a different genome:

1. Select **File > Open Genome or Array**.
2. Follow the submenus to select the genome you want.

Opening Another Copy of the Main Window

To open another copy of the main window:

- Select **File > New Linked Window**.

This command brings up a new main window similar to the one described in "GeneSpring Basics" on page 3-5.

Changing Preferences

To change preferences, such as set up a data directory, choose a color scheme, or configure a firewall proxy, choose **Edit > Preferences**. See "Setting Preferences" on page 3-15 for more details.

Using the Gene Inspector Window

To display the Gene Inspector window:

- Double-click a gene in the Genome Browser.

Figure 3-3 shows an example of the Gene Inspector.

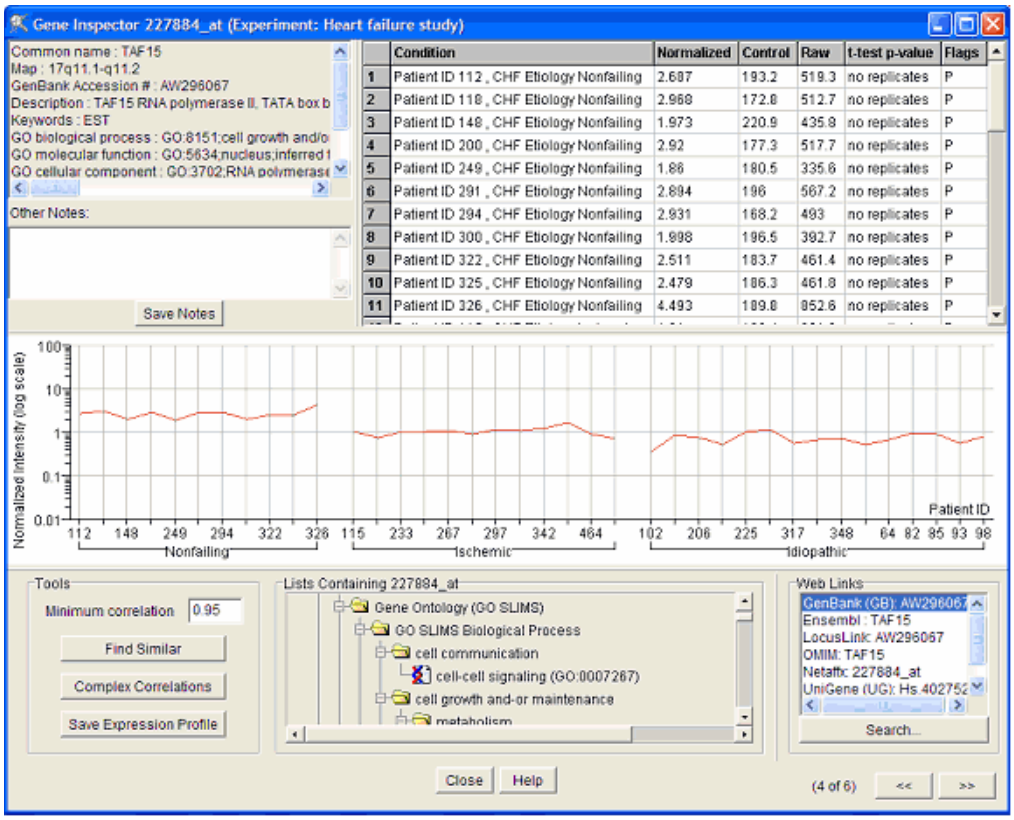

Gene Inspector 227884_at (Experiment: Heart failure study)
Common name : TAF15
Map : 17q11.1-q11.2
GenBank Accession # : AW296067
Description : TAF15 RNA polymerase II, TATA box b
Keywords : EST
GO biological process : GO:8151;cell growth and/o
GO molecular function : GO:5634;nucleus;inferred f
GO cellular component : GO:3702;RNA polymerase
Other Notes:
Save Notes
Condition
Normalized
Control
Raw
t-test p-value
Flags
1 Patient ID 112 , CHF Etiology Nonfailing 2.687 193.2 519.3 no replicates P
2 Patient ID 118 , CHF Etiology Nonfailing 2.968 172.8 512.7 no replicates P
3 Patient ID 148 , CHF Etiology Nonfailing 1.973 220.9 435.8 no replicates P
4 Patient ID 200 , CHF Etiology Nonfailing 2.92 177.3 517.7 no replicates P
5 Patient ID 249 , CHF Etiology Nonfailing 1.86 180.5 335.6 no replicates P
6 Patient ID 291 , CHF Etiology Nonfailing 2.894 196 567.2 no replicates P
7 Patient ID 294 , CHF Etiology Nonfailing 2.931 168.2 493 no replicates P
8 Patient ID 300 , CHF Etiology Nonfailing 1.998 196.5 392.7 no replicates P
9 Patient ID 322 , CHF Etiology Nonfailing 2.511 183.7 461.4 no replicates P
10 Patient ID 325 , CHF Etiology Nonfailing 2.479 186.3 461.8 no replicates P
11 Patient ID 326 , CHF Etiology Nonfailing 4.493 189.8 852.6 no replicates P
Normalized Intensity (log scale)
100
10
1
0.1
0.01
112 148 249 294 322 326 115 233 267 297 342 464 102 206 225 317 348 64 82 85 93 98
Patient ID
Nonfailing
Ischemic
Idiopathic
Tools
Minimum correlation 0.95
Find Similar
Complex Correlations
Save Expression Profile
Lists Containing 227884_at
Gene Ontology (GO SLIMS)
GO SLIMS Biological Process
cell communication
cell-cell signaling (GO:0007267)
cell growth and-or maintenance
Web Links
GenBank (GB): AW296067
Ensembl : TAF15
LocusLink: AW296067
OMIM: TAF15
Netaffx: 227884_at
UniGene (UG): Hs.402752
Search...
Close
Help
(4 of 6)
<<
>>

Figure 3-3 Gene Inspector Window

This window contains specific information about the selected gene. See "Using the Gene Inspector" on page 7-16 for details. Information presented in the Gene Inspector might include:

- Knowledge you have about your selected gene (typically text).
- Graphs of the selected gene's expression profile from the current experiment.
- Links to Internet or Intranet databases on the web for the selected gene.

Making Gene Lists

You can make gene lists from within the Gene Inspector window.

Making Lists with the Find Similar Command

The **Find Similar** button in the Gene Inspector lets you create a list of genes having similar expression profiles to the gene being displayed. See "Inspecting Gene Lists" on page 12-11 for more details.

Making Lists with the Complex Correlation Command

The **Complex Correlation** button in the Gene Inspector lets you make a list of all the genes satisfying various similarity measurements you apply. See "Correlation of Expression Change" on page 14-11 for more details.

Making Lists with the Venn Diagram

The **Colorbar > Color by Venn Diagram** command lets you make lists based on the membership of genes in a Venn Diagram. Right-clicking over lists in the Navigator lets you fill the diagram. See "Making Gene Lists from Venn Diagrams" on page 12-18 for more details.

Making Lists with the Filter Genes Command

The **Filtering > Filter on Gene List Numbers** command lets you use expression level constraints and control strength restrictions to create a smaller gene list. See "Filtering on Gene List Numbers" on page 9-27 for more details.

Making Lists from Selected Genes

You can make a list of all the genes you have selected in the Genome Browser by right-clicking and choosing **Make List from Selected Genes**.

See "Managing Window Elements" on page 8-1 for information on how to select genes. See "Making Gene Lists from Selected Genes" on page 12-21 for more details on this method of making a gene list.

Making Lists from Conjectured Regulatory Sequences

Once you have found possible regulatory sequences using the Find Potential Regulatory Sequences window and are inspecting one of the sequences in the Conjectured Regulatory Sequence window, you can make a list of all of the genes containing that sequence by selecting the **List > Make Gene List** command.

See "Finding Potential Regulatory Sequences" on page 14-1 and "Making Gene Lists from Conjectured Nucleotide Sequences" on page 14-9 for more information.

Setting Preferences

The Preferences window lets you change global preferences for GeneSpring. It provides access to nine tabs for modifying many default settings. Note that some changes may not take effect in the currently open window or in your current GeneSpring session. Saved changes in the preferences window will not take effect until GeneSpring is restarted.

Opening the Preferences Window

To open the Preferences window:

- Select **Edit > Preferences**.

To change any options in the Preferences window:

- Click the appropriate tab to view the available settings.

Data Files Tab

The Data Files tab (Figure 3-4) lets you set the defaults of what you would like to see when GeneSpring opens. Set the defaults on this tab to enable GeneSpring to open your chosen genome.

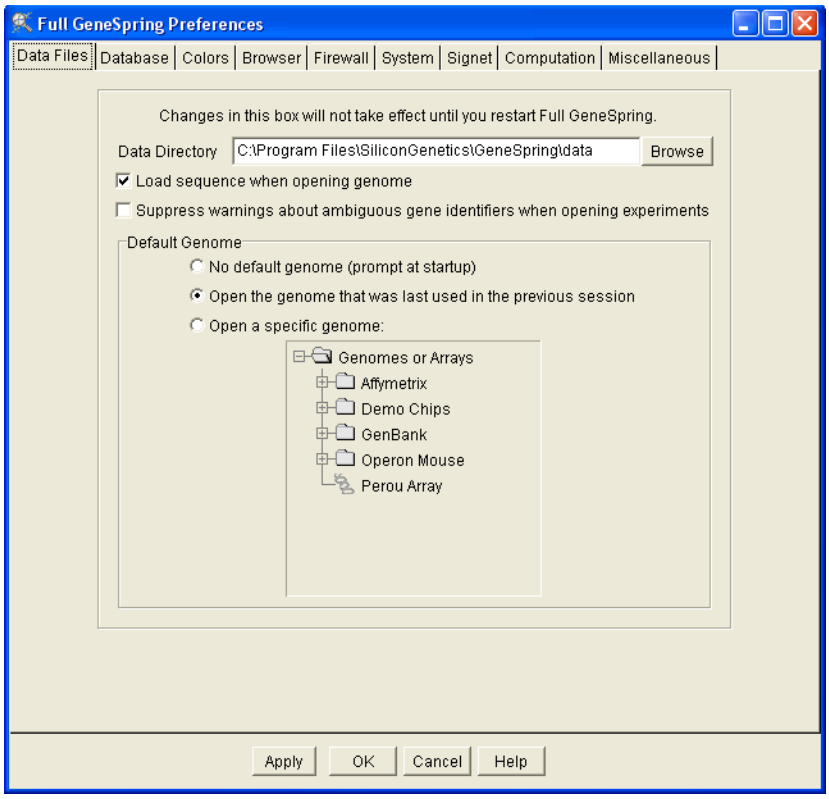


Figure 3-4 Preferences Window: Data Files Tab

- **Data Directory**—The directory containing all GeneSpring data, including the genome that opens at startup. Use the browse button or the Navigator to choose the directory.
- **Load Sequence**—Load nucleic acid sequences with the genome data at startup.
- **Suppress warnings about ambiguous gene identifiers when opening experiments**—Check this box to suppress the ambiguous gene identifier warning message (not recommended). For more information on ambiguous gene identifiers, see "Handling Ambiguous Gene Identifiers" on page 5-15.
- **Default Genome**—The default genome to open when you start GeneSpring.
  - Select **No default genome to be prompted for the genome** to open each time you start GeneSpring.
  - Select **Open the genome that was last used in the previous session to default** to the last genome opened.
  - Select **Open a specific genome to specify a default genome** to open every time you start GeneSpring. To change this value, select the desired genome from the displayed directory.

Note: On MacOsX, this menu is not displayed. To select a genome on MacOSx, click the **Browse** button.

Database Tab

The Database tab (Figure 3-5) lets you specify how GeneSpring assigns parameters for a series of numeric values in your database. You must also specify the fully-qualified class name of the driver in the JDBC driver field.

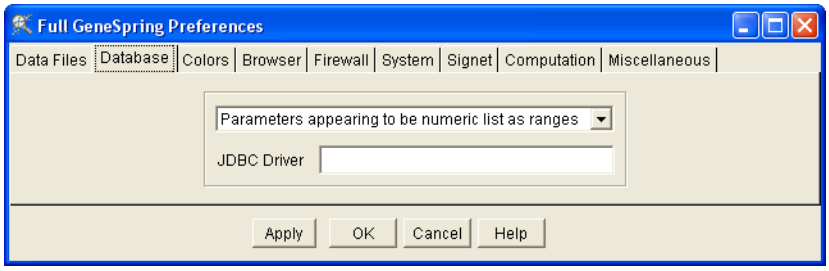


Figure 3-5 Preferences Window: Database Tab

Colors Tab

The Colors tab (Figure 3-6) lets you change the colors GeneSpring uses to represent different types of data and other window elements. There are many default color schemes to choose from. The brightness of a color depends on the trust associated with it. For more information, see "Trust" on page 8-61.

Over- and underexpressed color refers to the coloring of genes as shown in the Genome Browser and Colorbar. To change the definitions of overexpressed (upregulated) and underexpressed (downregulated) genes, right-click over the Colorbar in the main Genome Browser. See"Changing the Colorbar Range" on page 8-63 for more details on this topic.

The colors you choose are blended to create a continuous spectrum from High to Normal to Low expression values. There are two sections on this tab: Standard Colors and Group Colors.

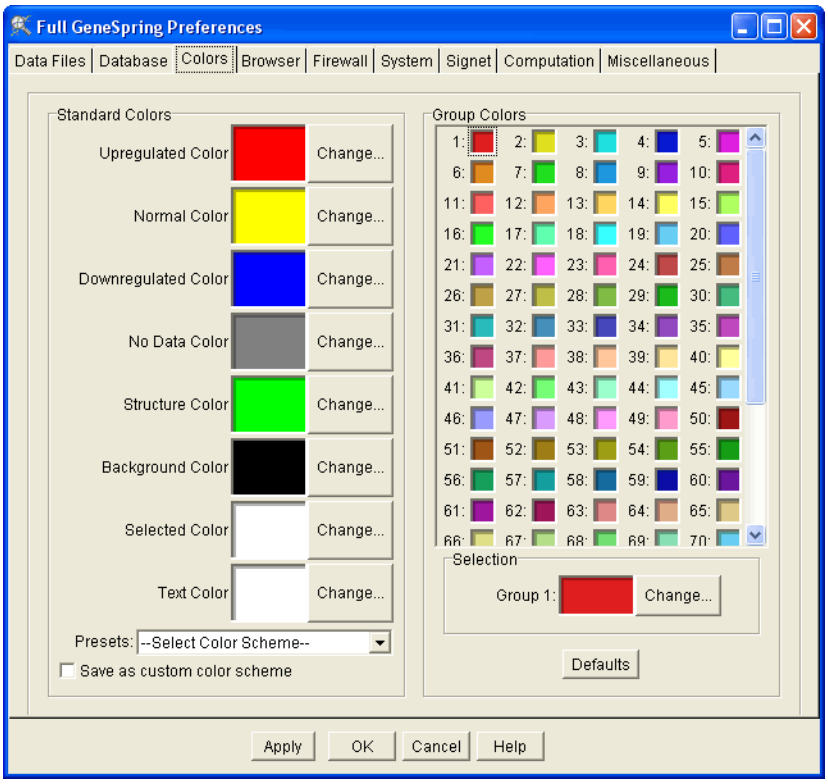


Figure 3-6 Preferences Window: Colors Tab

Standard Colors

The standard colors apply only to general display options.

- **Upregulated Color**—Used to display genes greater than or equal to the *High Expression* value selected for the current color bar.
- **Normal Color**—Used to represent genes having a normalized expression value of one. This is the only setting for which you can specify “no color”.
- **Downregulated Color**—Used to display genes less than, or equal to, the low expression value selected for the color bar.
- **No Data Color**—Used when expression data for a gene is not available in an experiment.
- **Structure Color**—Used for the Condition line and the lines between the genes in the Physical Position View, the Tree lines, the Ordered List lines, etc.
- **Background Color**—Defines the color behind the genes and other elements in the Genome Browser.
- **Selected Color**—Used for selected genes, gene names, and axes. For this, you will probably want the greatest contrast with the background color.
- **Text Color**—Defines the color of text displayed in the Genome Browser window.
- **Presets**—Lets you choose from a variety of pre-defined color schemes.

To create a custom color scheme, modify colors as desired and check the **Save as custom color scheme** box. This saves your current color scheme in the Presets menu under the name “Custom Color Scheme”. You can save only one custom color scheme at a time.

Group Colors

The Group Colors section lets you set colors used for the following:

- Classifications
- Parameters
- Gene Lists
- PCA
- Gene Inspector
- Find Similar Samples/Color by Attribute

Each box in the displayed grid indicates a color for that group.

Click on a box to select it. The Selection area at the bottom of the panel displays that box with the name and color of the selected group. Double-click the selected box or click **Change** to view the Change Color window. To restore the color defaults, click **Defaults**.

Specific Color Definition

You have the option to define your own colors to use in the Genome Browser. If your printer requires exact color definitions, specify them on this window. To change or adjust a color, select the **Change** button next to its element in the Color window (Figure 3-7).

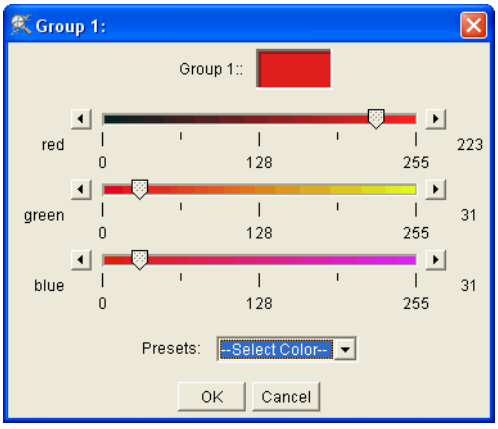


Figure 3-7 Preferences Window: Color Creation

Click over any slider and move it horizontally to adjust the color. Watch the color preview box and stop moving the cursor when the desired color is reached. Click **OK** to accept the new color. The **Specify no color** check box is only available for the "Normal Color" settings.

Browser Tab

The Browser tab (Figure 3-8) lets you specify default settings if you want to use a particular browser for the GeneSpring application. You only need to set the **Arguments** option if you are using an obscure browser that requires an argument.

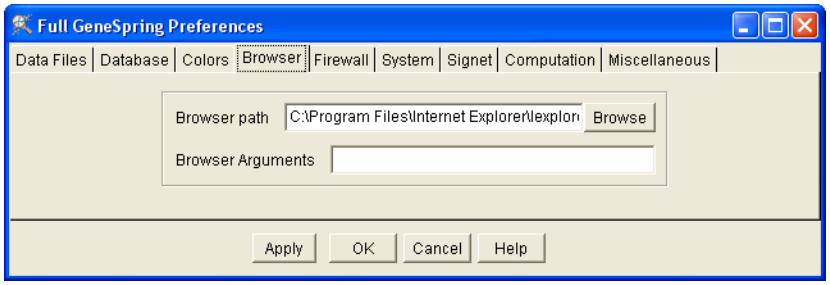


Figure 3-8 Preferences Window: Browser Tab

Firewall Tab

If your organization has a firewall, you may need to specify settings to allow GeneSpring to access outside networks (Figure 3-9).

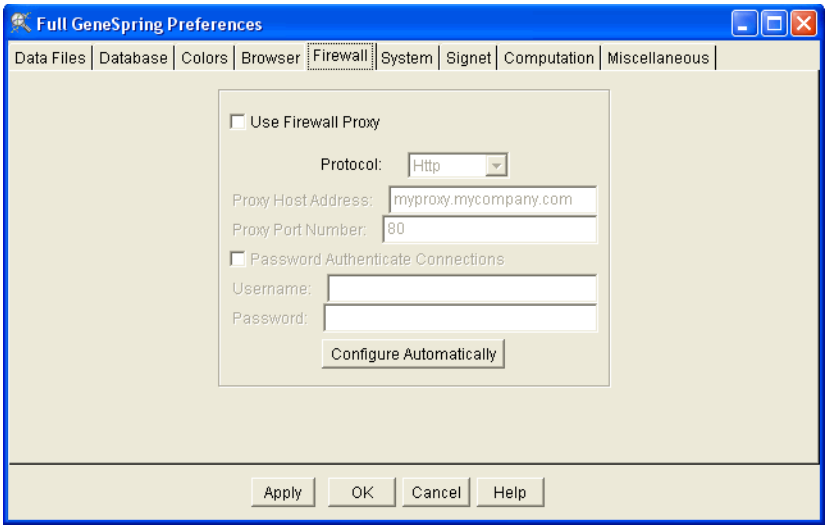


Figure 3-9 Preferences Window: Firewall Tab

Click **Configure Automatically** to have GeneSpring attempt to automatically detect the appropriate settings. If the settings GeneSpring chooses do not allow you to reach the Internet, you may need to alter these settings.

The following settings are available:

- **Protocol**—Specifies the firewall protocol to use. The options are: HTTP, SOCKS4, and SOCKS5.
- **Proxy Host Address**—Specifies the host address of the computer on which the firewall exists. This can be either a fully-qualified host name (for example, `hostname.domainname.com`) or an IP number.
- **Proxy Port Number**—Specifies the port on which to connect to the firewall host.
- **Password Authenticate Connections**—Specifies whether a password is required to connect to the firewall host.
- **User name**—Specifies the user name used to connect to the firewall host (if required).
- **Password**—Specifies the password used to connect to the firewall host (if required).

If you are unsure of how to proceed, contact your System Administrator for details about your firewall.

System Tab

The System tab (Figure 3-10) lets you specify a number of different parameters regarding networking and memory usage.

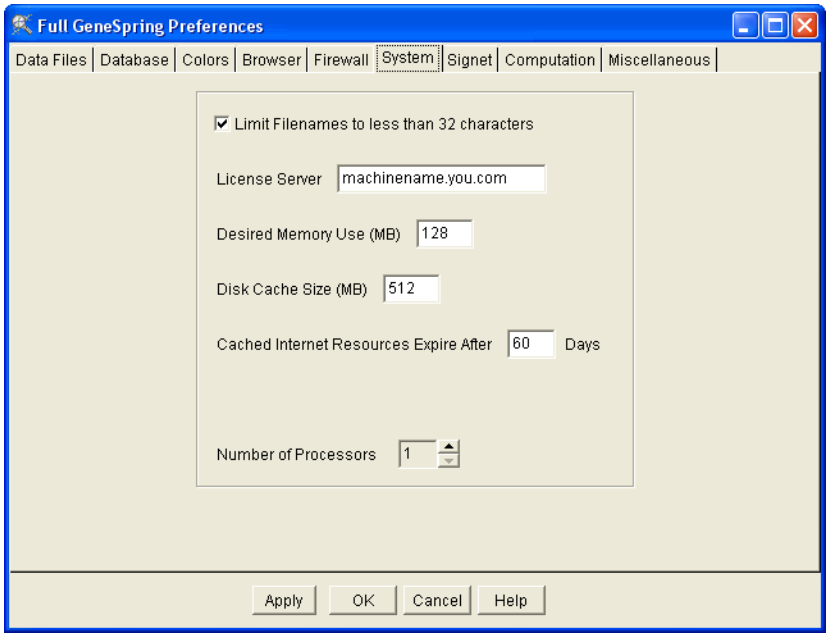


Figure 3-10 Preferences Window: System Tab

- **Limit Filenames to less than 32 characters**—Lets you limit the length of file names. This is a useful default setting for Macintosh users, since MacOS does not accept file names longer than 32 characters.
- **License Server**—Lets you specify the IP address of the machine that dispenses concurrent licenses.
- **Desired Memory Use**—Lets you set the amount of RAM GeneSpring attempts to use. If this value is set too high with respect to total available memory, unnecessary disk caching occurs and performance will be slow.
- **Disk Cache Size—**Specifies the amount of hard disk space GeneSpring uses to temporarily store HTML pages accessed by the GeneSpider or by other Internet-based search functions.

  Silicon Genetics recommends that you set this value to 10% of your available disk space. In GeneSpring, experimental data is loaded into the disk cache instead of into system RAM. GeneSpring now loads only the data currently being used into memory. This enables GeneSpring to handle much larger experiments containing any number of samples.
- **Cached Internet Resources Expire After**—Specifies how long GeneSpring caches copies of Internet resources for quicker access.
- **Number of Processors**—Specifies the number of processors in your computer. This allows several types of analysis (including k-means, build gene trees, promoter search, and class predication) to be used most efficiently.

Signet Tab

On the Signet tab (Figure 3-11), you can specify the default Signet server to use and enter the addresses of other Signet servers you might want GeneSpring to access.

- To enable GeneSpring to automatically connect to the default Signet server at startup, check the **Login to Signet at Startup** box. Select the default Signet server from the menu.
- To enable GeneSpring to invoke the Bulk Upload to Signet window when you quit GeneSpring, check the **Remind me to upload new data to Signet** box.
- When you create a new experiment in GeneSpring using samples from a Signet server, GeneSpring saves local copies of those samples by default. This can cause slow performance when saving large experiments. To disable this feature, clear the **Save Signet Samples Locally When Creating an Experiment** box.

To enter a new Signet server, click **New** and enter the following information:

- **Signet Server Name**—Specifies the name of the server to connect to
- **Signet Server Address**—Specifies the IP address of the Signet server. Enter the numeric address only. For example: `127.0.0.1`. Do not enter `“http://”` before this address.
- **Default User Name**—Specifies the user name to connect to Signet.
- **Signet & GeneSpring on**—Specifies whether the Signet server is on the same side of your firewall as GeneSpring or not.

- **Use Secure Connection**—Specifies whether communication between GeneSpring and Signet should be secured or not. If it is secured, communication between Signet and GeneSpring uses HTTPS (which uses the SSL library available in Java).

To edit an existing Signet server, select it from the list and click the **Edit** button. To delete a Signet Server, select it from the list, and then click the **Delete** button.

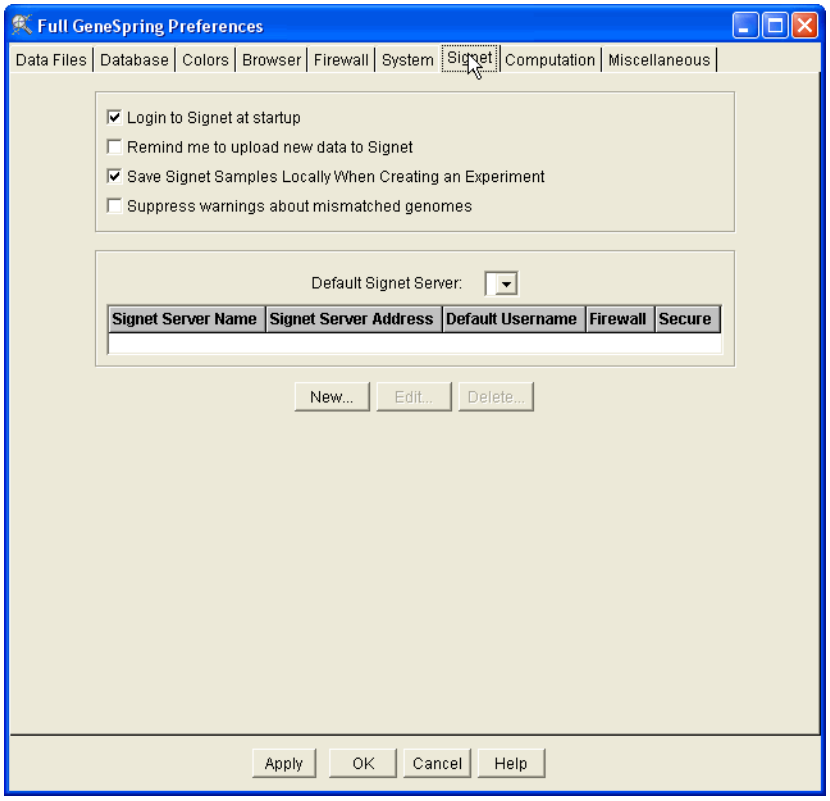


Figure 3-11 Preferences Window: Signet Tab

Computation Tab

The Computation tab (Figure 3-12) lets you specify settings for running scripts.

- **Default Computation**—Select **Local** to have scripts run on your local machine by default. Select Remote to run scripts on a remote execution server by default.
- **Local Computation Settings**—Select **Don't Show Script Result Summary Window** to skip the Script Result Summary when a script completes its execution. This option applies to scripts that produce fairly complex multiple group results.

  The **Current Scale Factor for Time Estimate** option lets you reset the multiplier for GeneSpring's internal estimate of how long an analysis will take back to the default value of 1. This can be useful if you have significantly changed hardware or settings on your computer (added more RAM, etc.).
- **Remote Computation Settings—**Select **Automatically check for results** to enable GeneSpring to automatically check whether your script has finished running. Specify how often to check by entering a number of minutes in the **Delay between checks** box.

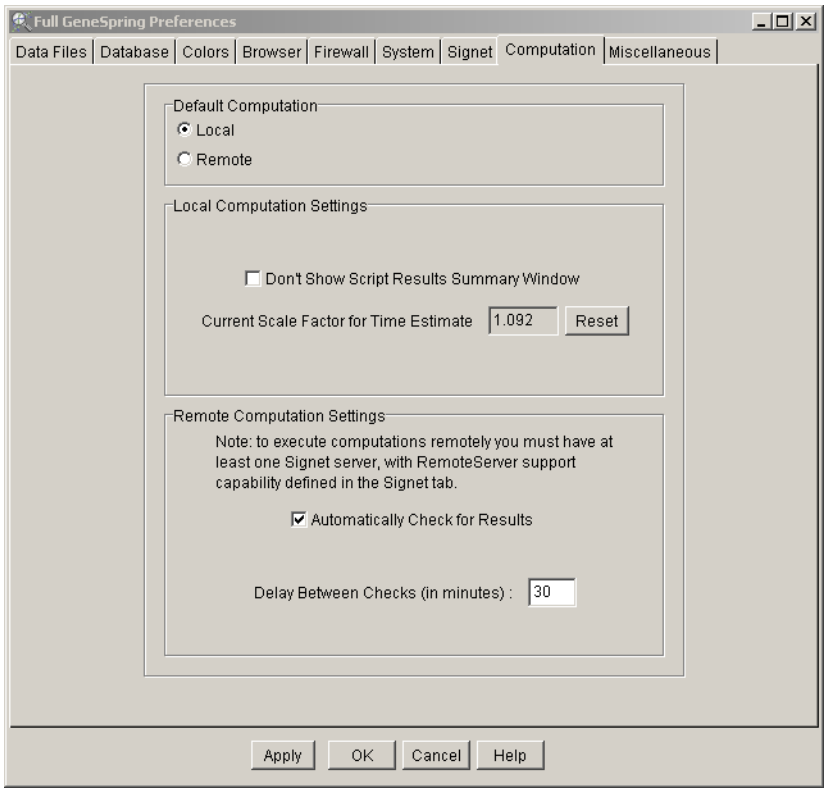


Figure 3-12 Preferences Window: Computation Tab

Miscellaneous Tab

The Miscellaneous tab (Figure 3-13) contains a variety of settings to customize your GeneSpring installation.

- **Default Minimum Correlation**—Specifies the default minimum correlation coefficient that appears near the Find Similar button in the Gene Inspector window.
- **Restrict Gene List Searches**—Limits the lists GeneSpring examines when searching for similar lists in the Gene List Inspector window. The options are:
  - **Standard lists**—Limits searches to only standard lists. For example, the GO SLIMS ontology gene list represents a special type of gene list that GeneSpring identifies as a standard list.

    To create a standard list, go to "Using the Gene List Inspector" on page 7-33 for more information.
  - **All Lists**—Enables searches in all lists.
  - **No Lists**—Disables the Similar Lists function.
- **Search Gene Lists Stored**—Specify whether to search gene lists stored on your local machine, on a Signet server, or both.

- **Use the Cross-Gene Error Model by Default in Experiment Interpretations**—Select this option to use the Cross-Gene Error Model by default in experiment interpretations.
- **Font Name and Size**—Specifies the style and point size of the default display font.
- **Default Font**—Click this button to reset the display font to its default value.
- **Language**—Lets you select from the available language choices for the GeneSpring interface. If your computer is set for a specific language, use the same setting here.
- **Your Name, Your Research Group, Your Email**—Specifies the name, group name, and email address values contained in the HTML files that go into your data directories.
- **Entrez mirror**—Specifies a web address for a mirrored site.

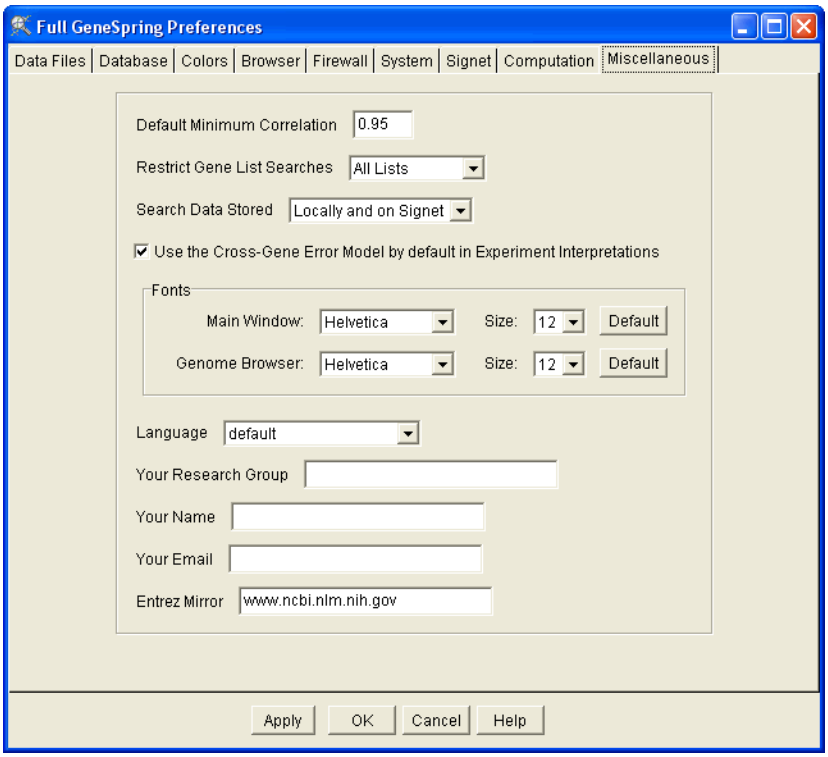


Figure 3-13 Preferences Window: Miscellaneous Tab

4

Importing Genomes

This chapter explains how to import genomes into GeneSpring. It covers the following topics:

- Overview
- Importing Genomes from Silicon Genetics
- Importing Genomes from GenBank/EMBL Files
- Importing Tab-delimited Genome Files
- Managing Web Links
- Saving New Genomes
- Updating Annotations with GeneSpider
- Making Gene Lists from Annotations
- Building Ontologies
- Importing KEGG Pathways
- Building Homology Tables
- Obtaining Supplemental Information

Overview

In the context of GeneSpring, a genome contains information about all the genes in your chip or microarray. Thus, the definition of a genome in GeneSpring does not correspond exactly to the biological definition of a genome.

A genome, also referred to as an array in GeneSpring, is composed of discrete genes. This means that a GeneSpring genome can contain two genes representing alternately spliced variants of a single gene; whereas, a true genome would include the DNA sequences for only one genome.

A GeneSpring genome corresponds to one mRNA species, the measure of gene expression assays. One gene in the biological definition can give rise to many different mRNA species, and each of these species are represented by one gene in GeneSpring. Setting up a genome is usually the first step in the data analysis workflow.

Import Options

The following options are available for importing genomes or arrays:

- Download pre-made genomes from the Silicon Genetics site
- Download files from the GenBank or EMBL FTP sites
- Create genomes from data in a tab-delimited file

The Import Genome wizard guides you through each step in the import process.

Note: GeneSpring also enables you to create genomes from expression data. Refer to Chapter 5, "Working With Experiments" for complete information.

Import New Genome Wizard

Once you have imported a genome, the Import New Genome wizard will guide you through the process of preparing your files for data analyses. These steps include the following:

- Create a list of annotations (what the scientific community knows about each gene, including information used for building ontologies)
- Create a list of gene hypertext links (URLs from which you can find more information about each gene from public databases
- Map information about where each gene appears on a given chromosome
- Add gene identifiers (accession numbers) to the genes from various public databases

Importing Genomes from Silicon Genetics

This section explains how to download genomes from the Silicon Genetics web site.

Pre-made Genomes

Silicon Genetics can provide you with genomes for the Affymetrix, Agilent, Clontech, and CodeLink platforms. Be sure to visit the Silicon Genetics web site for the latest list of pre-made genomes.If you are using the ABI technology, contact ABI directly to obtain genomes for use in GeneSpring.

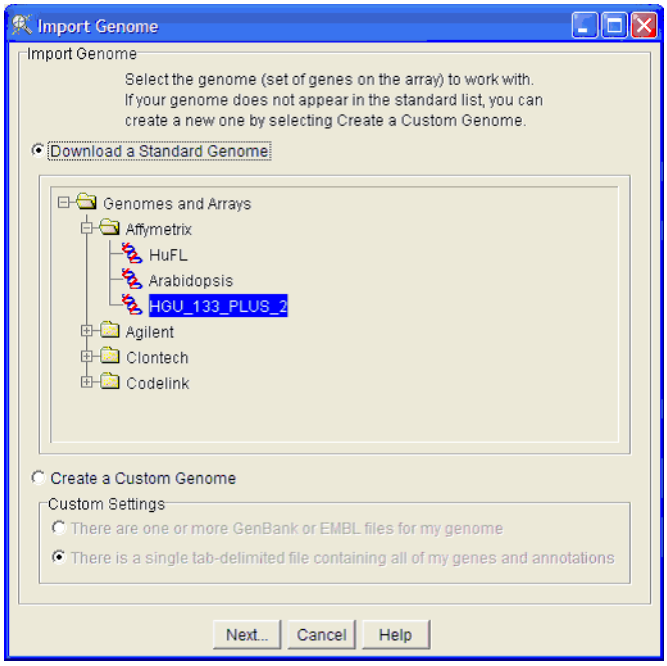
Importing Genomes from Silicon Genetics

To import new genomes from Silicon Genetics, you simply select the pre-made genomes you want and download them from the web site into GeneSpring.

To import a genome from Silicon Genetics:

1. Select **File > Import Genome**.

   The Import Genome window opens.


Import Genome
Import Genome
Select the genome (set of genes on the array) to work with.
If your genome does not appear in the standard list, you can
create a new one by selecting Create a Custom Genome.
Download a Standard Genome
Genomes and Arrays
Affymetrix
HuFL
Arabidopsis
HGU_133_PLUS_2
Agilent
Clontech
Codelink
Create a Custom Genome
Custom Settings
There are one or more GenBank or EMBL files for my genome
There is a single tab-delimited file containing all of my genes and annotations
Next...
Cancel
Help

2. Select the **Download a Standard Genome** option.

   The list of pre-made genomes is organized into folders and files. The folders list the name of the manufacturer and the folders contain the genomes from that manufacturer. Each genome is named with the exact name provided by the manufacturer. Different versions of a chip are treated as different genomes.

3. Single-click the **Genomes and Arrays** folder.

4. Navigate to the genome you want.

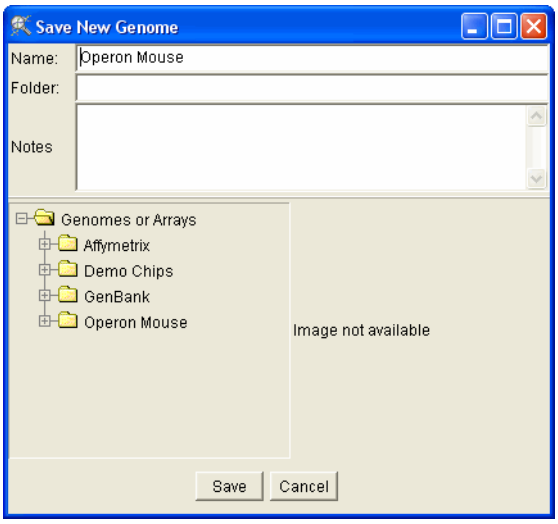
5. Click on a genome to select it.

6. Click the **Next** button.

   A progress bar opens and the genome begins to download. When the downloading has completed, the genome appears in a new Genome Browser. The import process is complete.

   Note: Not all technology providers have given Silicon Genetics permission to distribute the information about their arrays, since many consider this information proprietary information. If the genome or array you are using for your gene expression analysis is not listed in this window, we may be able to provide you with the genome. Please contact technical support at techsupport@silicongenetics.com or call us at +1-(866) SIG SOFT(866) 744 7638 (in the US) or +81 (70) 5072-2085 (Japan) or +44 (0) 1259 751 833 (Europe).

The Save New Genome window opens.



7. Go to "Saving New Genomes" on page 4-27 for instructions.

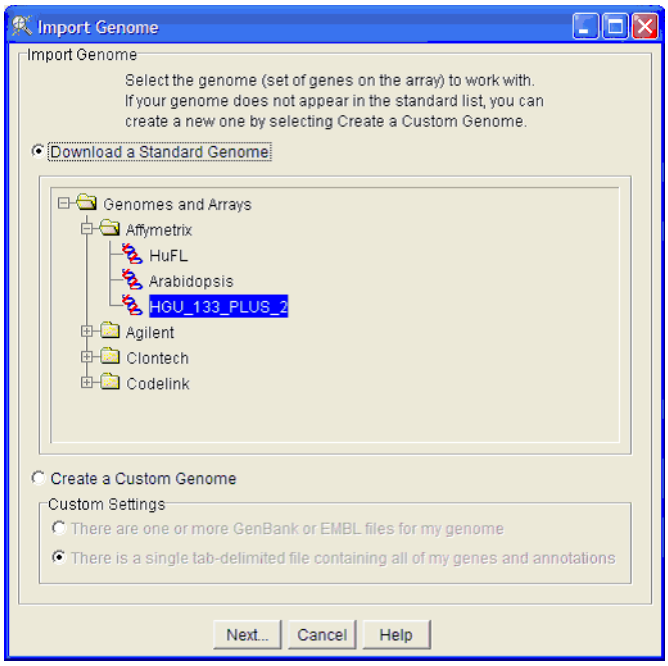
Importing Genomes from GenBank/EMBL Files

This section explains how to import genomes from the GenBank and EMBL (European Molecular Biology Laboratory) DNA databases through the GenBank (gbk) and EMBL (embl) file formats. To perform this procedure, you need access to one or more genomes from the genetic sequence database of your choice.

To import a genome from a GenBank/EMBL file:

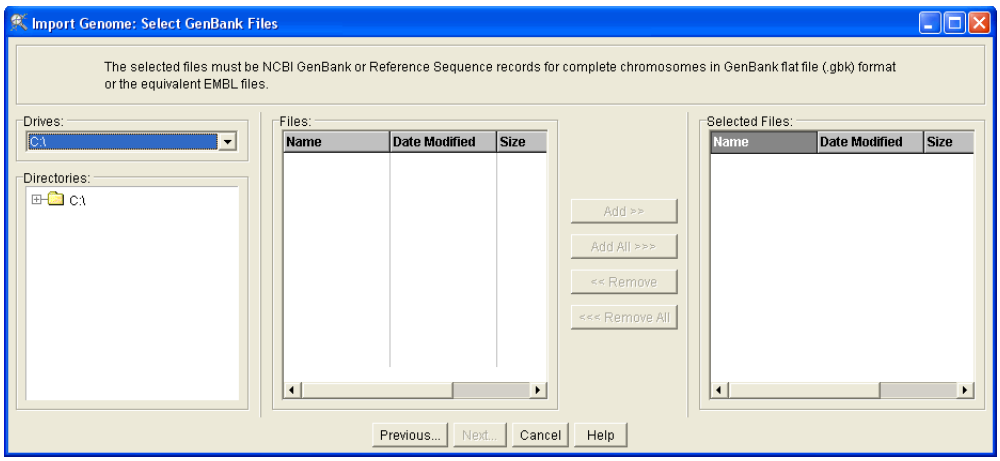
1. Select **File > Import Genome**.

   The Import Genome window opens.


Import Genome
Import Genome
Select the genome (set of genes on the array) to work with.
If your genome does not appear in the standard list, you can
create a new one by selecting Create a Custom Genome.
Download a Standard Genome
Genomes and Arrays
Affymetrix
HuFL
Arabidopsis
HGU_133_PLUS_2
Agilent
Clontech
Codelink
Create a Custom Genome
Custom Settings
There are one or more GenBank or EMBL files for my genome
There is a single tab-delimited file containing all of my genes and annotations
Next...
Cancel
Help

2. Select the **Create a Custom Genome** option.
3. Select the **There are one or more GenBank or EMBL files for my genome** option, if it is not already selected.
4. Click the **Next** button.

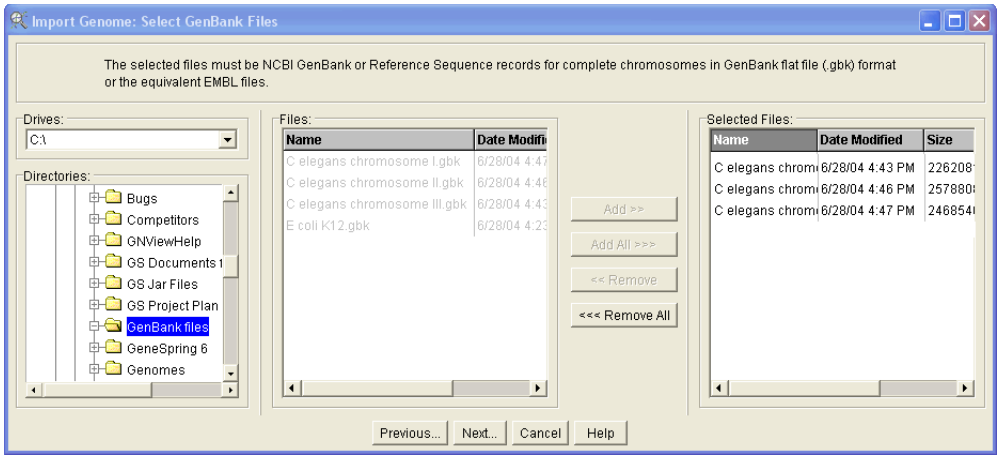
   The Import Genome: Select GenBank Files window opens.



5. In the **Drive** pull-down menu, select the drive you want.
6. In the **Directories Navigator**, select the directory you want.

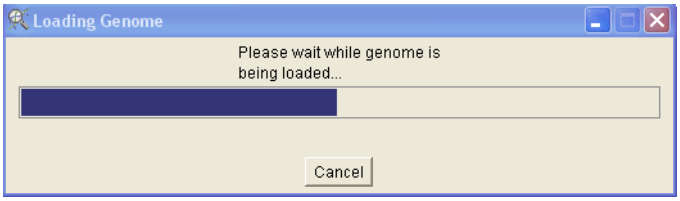
7. In the **Files Navigator**, do the following:
   - To sort by columns, click the column title.
   - To select one file, single-click the file.
   - To select several files at once, press Shift+click or Ctrl+click.
   - To deselect a file, press Ctrl+click.
8. Click the **Add**>> button to add the selected files, or click the **Add All**>> button to all the files.

   The selected files appear in the Selected Files column.



9. Click the **Next** button.

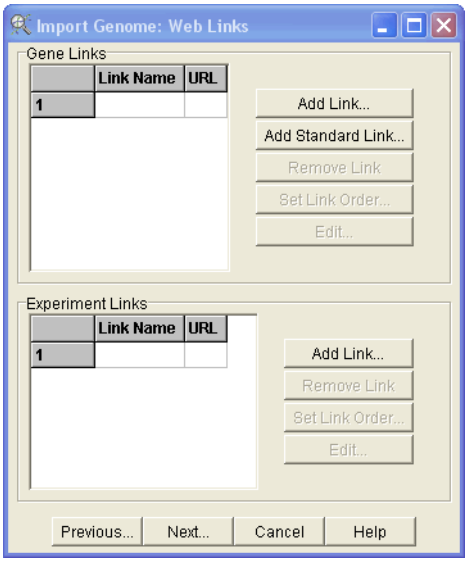
   The Loading Genome window opens.



10. Do one of the following:
    - If the Non-Unique Identifiers window opens, GeneSpring has determined that the systematic names of the genes within the genome are not unique.

      GeneSpring requires that the systematic names of genes are unique within a genome. When names in the gbk or embl files are not unique, GeneSpring will attempt to find a unique identifier within the entry to use. The results are presented in the Non-Unique Identifiers window.
    - If the Name Problems window opens, GeneSpring has detected a problem with long gene names. Go to "Fixing Long Gene Name Problems" on page 4-11 to resolve the problem.

- If the genome has loaded correctly, the Import Genome: Web Links window opens.



Go to "Managing Web Links" on page 4-15 for the next step.

Importing Tab-delimited Genome Files

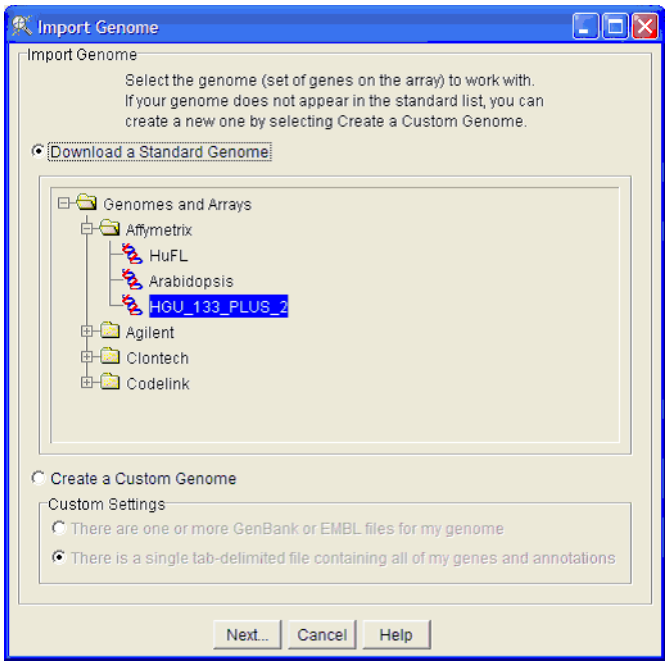
You can import genome files into GeneSpring if the file format is a tab-delimited text file. In GeneSpring, you need to specify the type of information that is included in the key columns of your data file. The minimal data file can contain just one column. This column contains the unique systematic names for each gene in the genome.

Selecting Annotation Files

To select an annotation file:

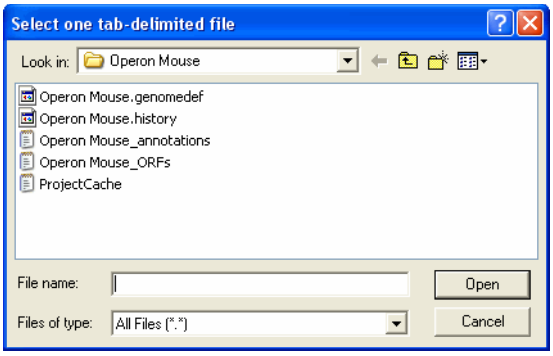
1. Select **File > Import Genome**.

   The Import Genome window opens.


Import Genome
Import Genome
Select the genome (set of genes on the array) to work with.
If your genome does not appear in the standard list, you can
create a new one by selecting Create a Custom Genome.
Download a Standard Genome
Genomes and Arrays
Affymetrix
HuFL
Arabidopsis
HGU_133_PLUS_2
Agilent
Clontech
Codelink
Create a Custom Genome
Custom Settings
There are one or more GenBank or EMBL files for my genome
There is a single tab-delimited file containing all of my genes and annotations
Next...
Cancel
Help

2. Select the **Create a Custom Genome** option.
3. Select the **There is a tab-delimited file containing all of my genes and annotations** option.

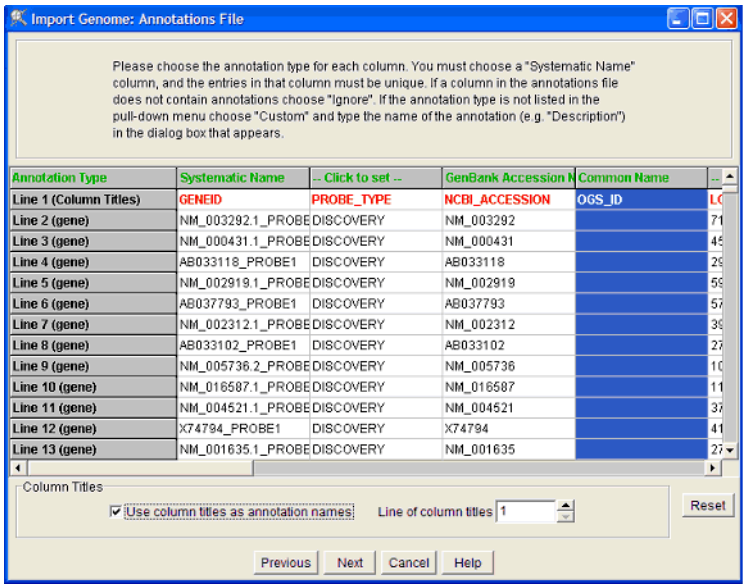
   The Import a Tab Delimited File window opens.



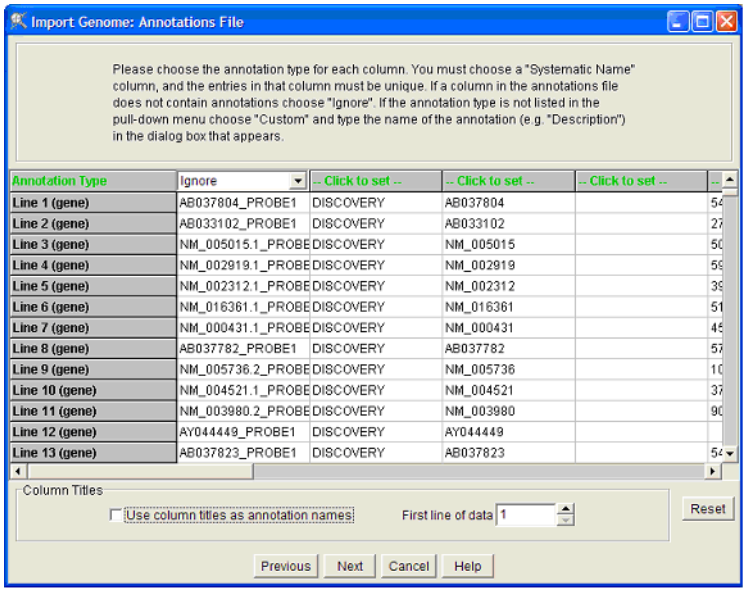
4. In the **Look in** menu, click the drive, folder, or Internet location that contains the file you want to open.
5. In the **Folder** menu, locate and open the folder that contains the file.
6. Select the file and click the **Open** button.

The Import Genome: Annotations File window opens. This window has two views:

- When the **Use column titles as annotation names** option is selected, the **Line of column titles** spinner box appears (as shown below).



- When the **Use column titles as annotation names** option is not selected, the F**irst line of data** spinner box appears (as shown below).



This window lets you choose the annotation type for each column. You must choose one column and label it as “Systematic Name”. The available annotation types are:

- **Systematic Name**—The unique identifier for the gene in this genome or array. It is recommended that the gene's systematic name be used to label the gene’s expression values in your experiment data files.
- **Common Name**—An alternative way of referring to this gene. Genes are not required to have a common name, and common names do not have to be unique. Using duplicate common names; however, will cause data not to be imported. Usually, the common name annotation column can be used to store the HUGO gene symbol or some other official gene identifier.
- **GenBank Accession Number**—The GenBank or EMBL identifier for this gene, if known. If the GenBank identifiers for your genes were not used as either their systematic or common names, then including the GenBank Accession Number in this field allows you to update the information about this particular gene directly from GenBank. See Updating your Master Gene Table with GeneSpider for more information.
- **Synonym**—This column allows for other names to be entered for the genes. Multiple names should be separated by semicolons (;).
- **Description**—A description of this gene, if known. This information can be accessed when you use the Find Gene command.
- **Map**—Mapping information for this gene. This should be a chromosome or a nucleotide position (1:228836..229309), or a cytogenic map position (such as 16q12.1).
- **Use column titles as annotation names**—Apart from the standard annotation columns that are described above, each gene can have many more annotation columns. Each annotation column needs to have a name and this name can be extracted from the tab-delimited file.
  - **Line of column titles**—This option only appears if **Use column titles as annotations names** is selected. In most cases, the first line of the file contains the titles for each of the columns. When you select this option, GeneSpring uses the line indicated by the “Line of columns titles” value in the menu. If the column header row contains titles that are not in the first row, enter the row number in which the titles appear.
  - **First line of data**—This option only appears if **Use column titles as annotations names** is not selected. Use this setting if either the data file does not contain a header line or when the headers are not appropriate.
- **Reset**—Resets the menus to **Click to Set**.

7. To set a **Systematic Name** column, do the following:
   a. Locate the column you want to label “Systematic Name.”
   b. From the menu, select the **Systematic Name** option.
8. If a column in the annotations file is blank or if you do not want to import the annotation column, leave the menu on the **Click to Set** option.

9. If you want an annotation that isn't in the menu, do the following:

   a. Select the **Custom** option from the menu.

      The Custom Annotation window opens.

   b. Enter the name you want and click the **OK** button.

      The name appears as the column header.

10. Click the **Next** button.

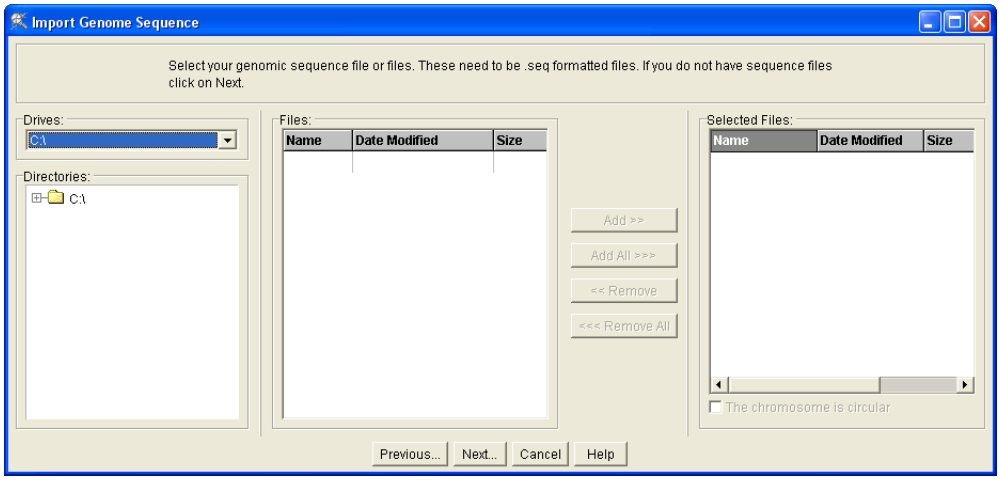
11. Do one of the following:

- If the Non-Unique Identifiers window opens, GeneSpring has determined that the systematic names of the genes within the genome are not unique.

  GeneSpring requires that the systematic names of genes are unique within a genome. When names in the gbk or embl files are not unique, GeneSpring will attempt to find a unique identifier within the entry to use. The results are presented in the Non-Unique Identifiers window.

- If the Name Problems window opens, GeneSpring has detected a problem with long gene names. Go to "Fixing Long Gene Name Problems" on page 4-11 to resolve the problem.

- If the Import Genomic Sequence window opens, you have successfully imported the tab-delimited file.



Go to "Importing Genome Sequences" on page 4-12 for the next step.

Fixing Long Gene Name Problems

The Name Problems window (Figure 4-1) indicates that GeneSpring has detected a problem with long gene names. The limit for name in the Systematic Names, Common names, GenBank accession numbers, and Synonyms columns of the Import Gene: Annotation File window is 256 characters. Long gene names are not allowed and must be resolved before you can proceed.

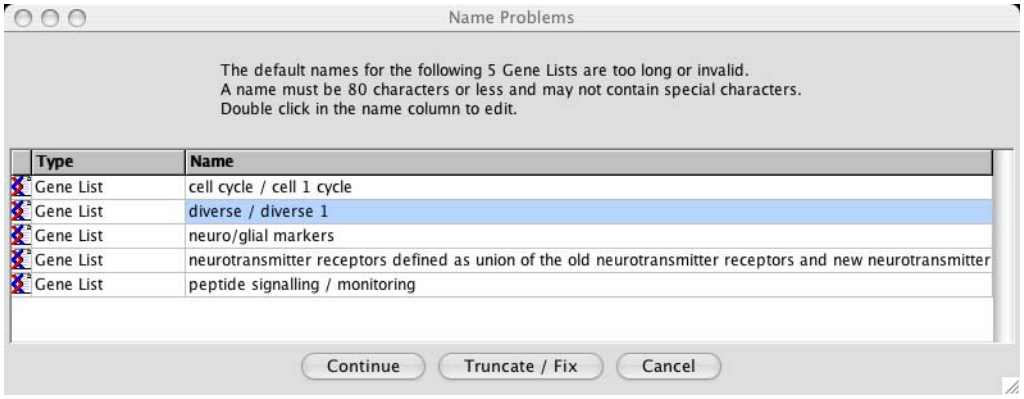

Name Problems
The default names for the following 5 Gene Lists are too long or invalid.
A name must be 80 characters or less and may not contain special characters.
Double click in the name column to edit.
Type
Name
Gene List
cell cycle / cell 1 cycle
Gene List
diverse / diverse 1
Gene List
neuro/glial markers
Gene List
neurotransmitter receptors defined as union of the old neurotransmitter receptors and new neurotransmitter
Gene List
peptide signalling / monitoring
Continue
Truncate / Fix
Cancel

Figure 4-1 Name Problems Window

This window contains the following elements:

- **Table**—Displays the results that have name problems.
- **Continue**—Disabled until the problems are fixed.
- **Truncate/Fix**—Truncates or fixes all of the problem entries based on their allowable size and characters.
- **Cancel**—Cancels the operation and closes the window.

To fix long names or invalid characters:

1. Do one of the following:
   - Manually fix all of the problem entries by editing them.
   - Automatically fix all of the problem entries by clicking the **Truncate/Fix** button.
2. Click the **Continue** button to fix the problem.

   The Import Genomic Sequence window opens.
3. Go to "Importing Genome Sequences" on page 4-12 for instructions.

Importing Genome Sequences

This section explains how to import custom genome sequence files in GeneSpring. The files that you import will be added to a Sequence Files table. The data must be in seq file format. GeneSpring will attempt to parse the chromosome names in the sequence files automatically. If the chromosome names cannot be parsed, GeneSpring will use the default chromosome names (numbers).

Sequence File Format

GeneSpring loads in sequence data from a GenBank or EMBL files automatically. If you have sequence data that is not in a GenBank/EMBL file, place it in a separate file using the seq format.

The Silicon Genetics seq format is similar to the FASTA format, although there are some differences. A FASTA formatted file, however, can easily be changed to a seq file. It basically only requires changing the identifier from the FASTA file to the chromosome number.

The seq format consists of one line of identifiers followed by lines of sequence. The identifier line consists of the “Greater than” sign (>) followed by the chromosome identifier, followed by a space which is followed by an optional description. An example is given here.

```
>CHR1 This is the description of Chromosome 1
GCTGACGGACTTTCTAGCGGTCTAGCAACTGAGCGGCGCGCGGGCATCGTA
CAGCAGCGAGCTACTATCTACGCGCGGCGGATATAAAACTACAAAAAAAAA
```

Chromosomes in GeneSpring are given a number (1, 2, 3 etc.) and the number should be part of the chromosome identifier. The chromosome identifier can optionally contain the letters 'CHR' but is not required. The number used in the seq format for the chromosome has to correspond to the number used in the Map position in the Master Table of Genes.

The seq format is not the same as the FASTA format. There is an example of the FASTA format at http://www.ncbi.nlm.nih.gov/BLAST/fasta.html.

An abridged example of the *yeast.seq* file might look like this:

```
>CHR1 Chromosome I data:
CCACACCACACCCACACACCCACACACCACCACCACACCACACCCACACACACA . . .
GTGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGGG
>CHR2 Complete DNA sequence of yeast chromosome II.
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGA . . .
AGAATAGGGTACTGTTAGGATTGTGTTAGGGTGTGGGTGTGGTGTGTGTGGG
TGTGGTGTGTGGGTGTGT
>CHR3 LOCUS       SCCHRIII   315341 bp    DNA             PLN
25-NOV-1996
CCCACACACCACACCCACACCACACCCACACACCACACACACCACACCCA . . .
AGTGTGTGGGTGTGGGTGTGTGGGTGTGGTGTGTGGGTGTGGTGTGTGTGTGGTGT
GTGGGTGTGGGTGTGTGGGTGTGGTGGGTGTGGTGTGTGTG
```

Name multiple chromosomes sequentially, for example, CHR1, CHR2 and so on. If there is only one chromosome, name it CHR1.

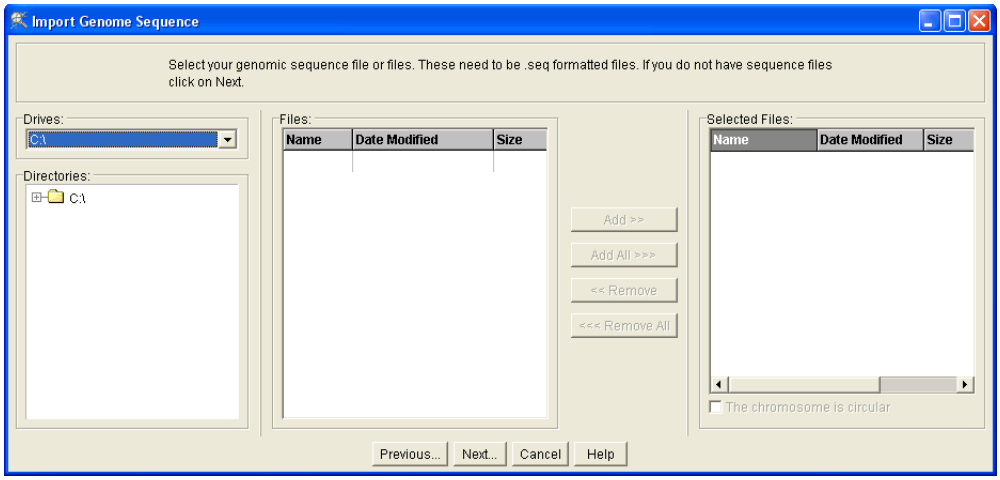
Importing Genome Sequences

This section explains how to import genome sequence files in seq format.

To import a genome sequence:

1. In the **Import Genome Sequence** window, do the following:
   a. In the **Drive** menu, select the drive you want.

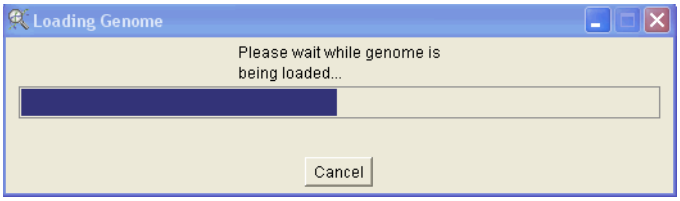
b. In the **Directories Navigator**, select the directory you want.



2. In the **Files Navigator**, do the following:
   - To sort by columns, click the column title.
   - To select one file, single-click the file.
   - To select several files at once, press Shift+click or Ctrl+click.
   - To deselect a file, press Ctrl+click.
3. Click the **Add>>** button to add the selected files, or click the **Add All>>** button to all the files.

   The selected files appear in the Sequence Files box.
4. If there is a single chromosome in the genome, and it is circular, select the **The chromosome is circular** option.
5. Click the **Next** button.

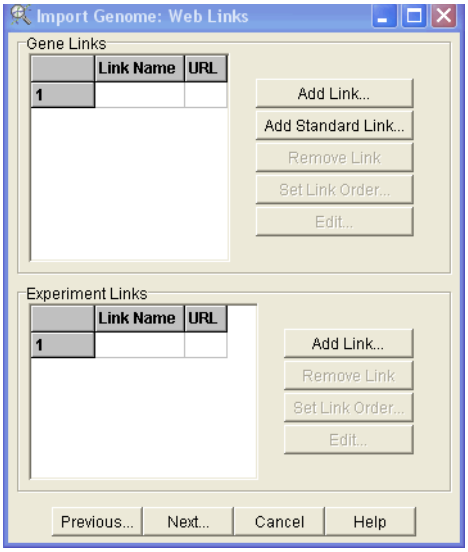
   The Loading Genome window opens.



6. Do one of the following:
   - If the Non-Unique Identifiers window opens, GeneSpring has determined that the systematic names of the genes within the genome are not unique.

     GeneSpring requires that the systematic names of genes are unique within a genome. When names in the gbk or embl files are not unique, GeneSpring will attempt to find a unique identifier within the entry to use. The results are presented in the Non-Unique Identifiers window.

- If the Name Problems window opens, GeneSpring has detected a problem with long gene names. Go to "Fixing Long Gene Name Problems" on page 4-11 to resolve the problem.
- If the genome has loaded correctly, the Import Genome: Web Links window opens.



Go to "Managing Web Links" on page 4-15 for the next step.

Managing Web Links

The Import Genome: Web Links window (Figure 4-2) lets you define the gene links and experiment links for the selected genome.

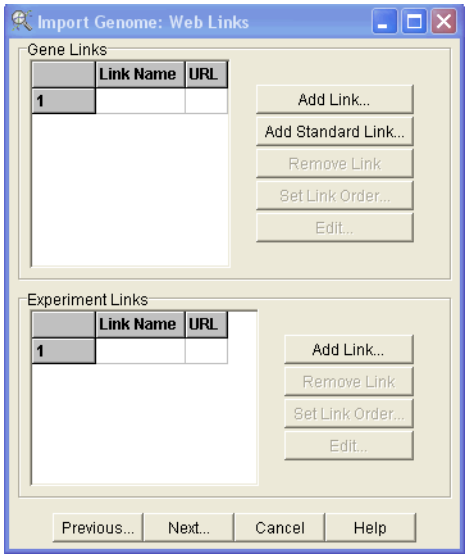


Figure 4-2 Import Genome: Web Links Window

Gene Links Table

The Gene Links table lets you define the web links that are available in the Gene Inspector window for this genome. It contains the following elements:

- **Link**—Displays the name GeneSpring will use to refer to a particular web link.
- **URL**—Displays the name of the web site GeneSpring will open when users select this URL.
- **Add Link Button**—Displays the Add Link window where you can type in a gene link.
- **Add Standard Link Button**—Displays the Add Standard Link window where you can select gene links from a list.
- **Remove Link Button**—Deletes the selected row.
- **Set Link Order Button**—Displays the Set Link Order window.
- **Edit Button**—Displays the Edit Link window. This option is only enabled when a row is selected.

Experiment Links Table

The Experiment Links table lets you define the web links that are available in the Condition Inspector window for this genome. It contains the following elements:

- **Link**—Displays the name of the web link.
- **URL**—Displays the URL for the web link.
- **Add Link Button**—Displays the Add Link window where you can type in a web link.
- **Remove Link Button**—Deletes the selected row.
- **Set Link Order Button**—Displays the Set Link Order window.
- **Edit Button**—Displays the Edit Link window. This option is only enabled when a row is selected.

For information on how to manage gene links, go to "Managing Gene Links" on page 4-16. For information on how to manage experiment links, go to "Managing Experiment Links" on page 4-22.

Standard Gene URLs

GeneSpring provides a number of standard gene URLs that you can add to the genome with the Add Standard Link button. These links can then be accessed through the Gene Inspector window.

Managing Gene Links

This section explains how to manage gene links in the Import Genome: Web Links window using the Add Link, Add Standard Link, Remove Link, Set Link Order, and Edit buttons. It covers the following topics:

- About Gene Links

- Adding Custom Gene Links
- Adding Standard Gene Links
- Setting the Gene Link Order
- Removing Gene Links
- Editing Gene Links

About Gene Links

GeneSpring 7 now has the ability to use different terms from the same annotation entry, provided that these terms are separated in some way (comma separated lists and semicolon separated lists are the most common). All current links will continue to work as they did in previous releases.

To enable this feature, GeneSpring has extended the hypertext link format using Perl5 compatible regular expressions, as follows:

- To cycle though a semicolon separated list of Common Names use:

  `<Common:([^;]*)>`

- To cycle though a comma separated list of Synonyms use:

  `<Synonym:([^,]*)>`

where `Synonym` indicates the name of the annotation and, `,` indicates the separation character. More complex regular expression can be constructed using the full power of regular expressions. It is also possible to include multiple columns in the construction of the regular expressions.

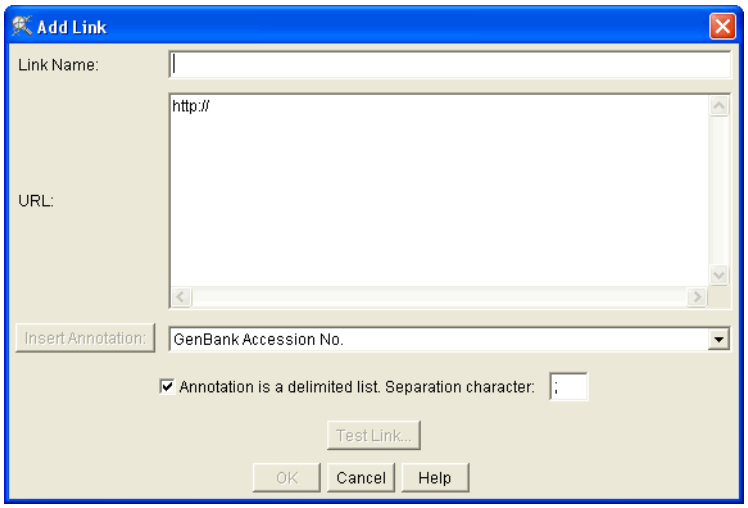
Adding Custom Gene Links

To provide users with access to additional information about a genome, you can use the Import Genome window to add a custom gene link. After you add the link, it will appear in the Gene Inspector window. Users who click the link will be redirected to the web page for the URL.

To add a custom gene link:

1. In the **Import Genome: Web Links** window, click the **Add Link** button in the Gene Links table.

   The Add Link window opens.



2. In the **Link Name** field, enter the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Gene Inspector.
3. In the **URL** field, paste the URL you want for a particular web database.

   For example:

   ```
   http://www.google.com/search?sourceid=
   navclient&q=<COMMON>:([^:]*)>+human+OR+sapiens
   ```

   will use the common name to query Google.
4. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Annotation** menu, select the annotation you want to add.
   c. To ensure that the annotation includes the regular expression for "loop over a delimited list" (where the delimiter is the character given in the Separation Character box), select the **Annotation is a delimited list** check box.

      The Separation Character box defaults to a semicolon.
5. Click the **Insert Annotation** button.

   The annotation is placed in the URL field with the correct punctuation. The default selection is the GenBank Accession Number or the Systematic Name (if the GenBank Accession Number is unknown).
6. To test the link, click the **Test Link** button.

   If the test fails, GeneSpring will be unable to open the link. Be sure you have entered a valid link.
7. Click the **OK** button.

   The Import Genome: Web Links window opens.

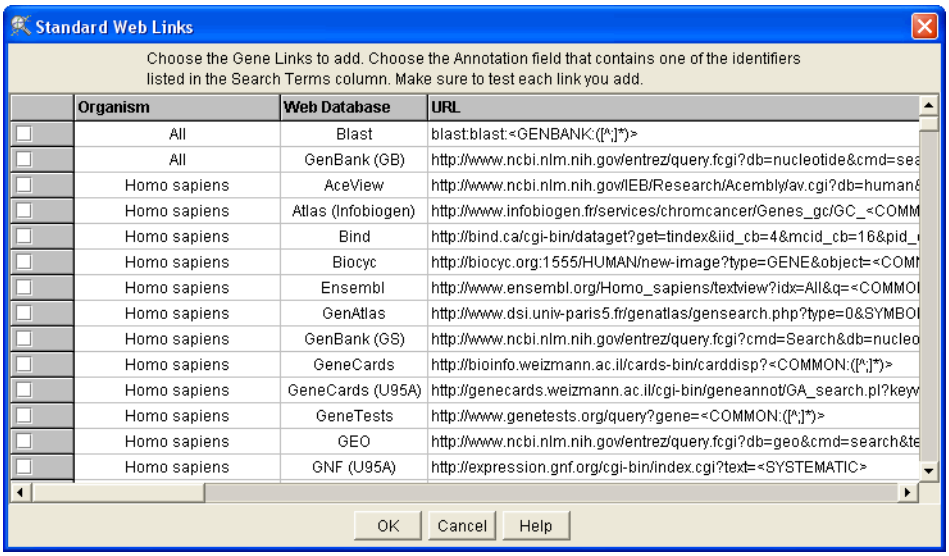
8. Do one of the following:
   - To add more web links, return to “Managing Gene Links” on page 4-16.
   - If you are done adding web links, go to “Saving New Genomes” on page 4-27.

Adding Standard Gene Links

To add a standard gene link:

1. In the **Import Genome: Web Links** window, click the **Add Standard Link** button in the Gene Links table.

   The Standard Web Links window opens.



   This window lets you easily import pre-made web links. It contains the following elements:

   - **Organism**—Displays the Latin names for the organism, or the term, “Various,” if the link can be applied to multiple organisms. For other Latin names, go to:
     - http://genome-www5.stanford.edu/cgi-bin/SMD/listMicroArrayData.pl?tableName=organism
     - http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi?view=platforms
   - **Web Database**—Displays the names of the links as they will appear on buttons in the Gene Inspector window.
   - **Column Headers**—Clicking on a column header sorts the table based on the entries.
   - **URL**—Displays the URL for the web link.

     For a partial list of standard gene links that you can add, go to “Standard Gene URLs” on page 4-16.
   - **Search Terms**—Provides examples of search terms that can be used in the Annotation list.

2. In the **Organism** column, locate the gene links you want to add.
3. Select the check box next to the links you want.

   Each check box you select will add a web link to the gene links table in the Import Genome: Web Links window.
4. In the **Annotation** list, select the annotation you want to use. The annotation should contain one of the types of annotations listed in the Search Terms column.

   Note: If you don't see the Annotation list, use the scroll bar to scroll to the right side of the window.
5. Click the **OK** button.

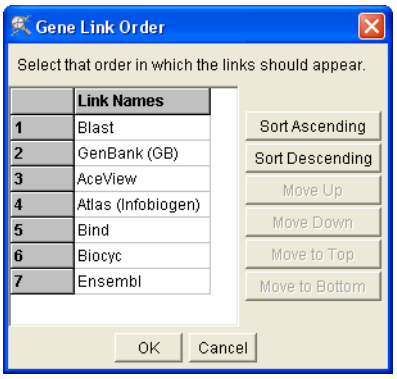
   The Import Genome: Web Links window opens.
6. Do one of the following:
   - To add more web links, return to "Managing Gene Links" on page 4-16.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Setting the Gene Link Order

To set the gene link order:

1. In the **Import Genome: Web Links** window, click the **Set Link Order** button in the Gene Links table.

   The Gene Link Order window opens.



   This window lets you set the order in which the links will appear in the Gene Inspector.
2. Do the following:
   - To sort the links in ascending order, click the **Sort Ascending** button.
   - To sort the links in descending order, click the **Sort Descending** button.
   - To move the selected link up one row, click the **Move Up** button.
   - To move the selected link down one row, click the **Move Down** button.
   - To move the selected link to the top of the column, click the **Move to Top** button.

- To move the selected link to the bottom of the column, click the **Move to Bottom** button.

3. Click the **OK** button.

   The Import Genome: Web Links window opens.

4. Do one of the following:

   - To add more web links, return to "Managing Gene Links" on page 4-16.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Removing Gene Links

To remove a gene link:

1. In the **Import Genome: Web Links** window, select a link to remove in the Gene Links table.
2. Click the **Remove Link** button.
3. Do one of the following:

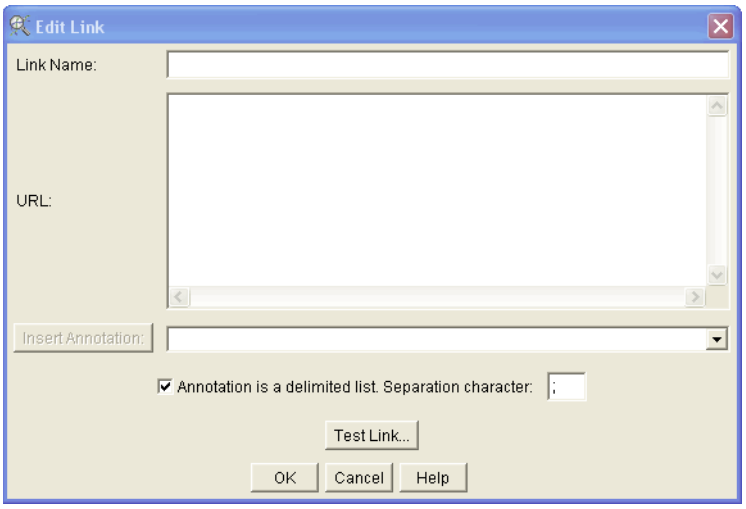
   - To add more web links, return to "Managing Gene Links" on page 4-16.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Editing Gene Links

To edit a gene link:

1. In the **Import Genome: Web Links** window, select a link to edit in the Gene Links table.
2. Click the **Edit** button.

   The Edit Link window opens.



3. In the **Link Name** field, edit the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Gene Inspector.

4. In the **URL** field, edit the URL you want for a particular web database.
5. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Annotation** menu, select the annotation you want to add.
   c. To ensure that the annotation includes the regular expression for "loop over a delimited list" (where the delimiter is the character given in the Separation Character box), select the **Annotation is a delimited list** check box.

      The Separation Character box defaults to a semicolon.
6. Click the **Insert Annotation** button.

   The annotation is placed in the URL field with the correct punctuation. The default selection is the GenBank Accession Number or the Systematic Name (if the GenBank Accession Number is unknown).
7. To test the link, click the **Test Link** button.

   If the test fails, GeneSpring will be unable to open the link. Ensure that the link you enter is valid.
8. Click the **OK** button.

   The Import Genome: Web Links window opens.
9. Do one of the following:
   - To add more web links, return to "Managing Gene Links" on page 4-16.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Managing Experiment Links

This section explains how to manage experiment links in the Import Genome: Web Links window using the Add Link, Remove Link, Set Link Order, and Edit buttons. It covers the following topics:

- About Experiment Links
- Adding Experiment Links
- Setting the Experiment Link Order
- Removing Experiment Links
- Editing Experiment Links

About Experiment Links

GeneSpring now has the ability to use different terms from the same annotation entry, provided that these terms are separated in some way (comma separated lists and semicolon separated lists are the most common). All current links will continue to work as they did in previous releases.

To enable this feature, GeneSpring has extended the hypertext link format as follows:

- To cycle though a semicolon separated list of Common Names use:

  `<Common:([^;]*)>`

- To cycle though a comma separated list of Synonyms use:

  `<Synonym:([^,]*)>`

where `Synonym` indicates the name of the annotation and the comma (`,`) indicates the separation character.

Adding Experiment Links

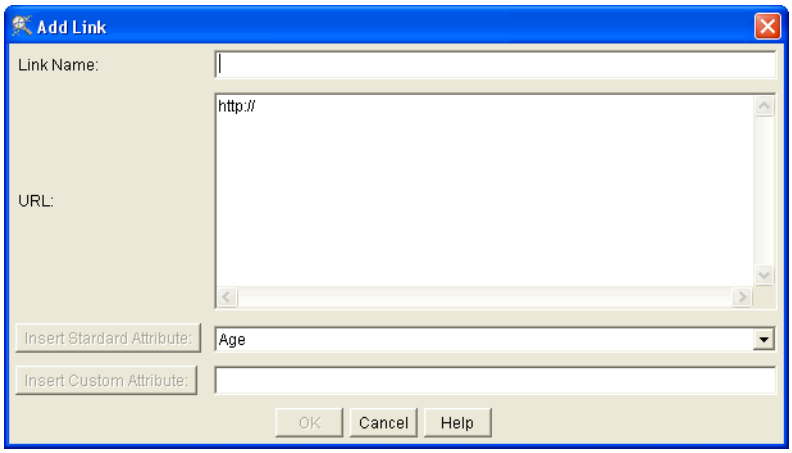
Experiment links allow you to link from a sample to some web site that contains more information about a sample or hybridization. Both sample attribute and parameter values can be used as the identifiers for the links to the web sites.

For example, to obtain more information about a disease that may have afflicted the donor of a sample, experiment links can be used to link to OMIM from the sample attribute called “Disease”. Similarly, experiment links can let a user link to an internal patient database that contains information about a patient based on the sample attribute, “Patient ID”.

To add an experiment link:

1. In the **Import Genome: Web Links** window, click the **Add Link** button in the Experiment Links table.

   The Add Link window opens.



2. In the **Link Name** field, enter the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Condition Inspector.

3. In the **URL** field, type or paste the URL you want for a particular web database.

   For example:

   `http://www.ncbi.nih.org.gov/geo/query/acc.cgi?acc=<GEO IDENTIFIER>`

will query the GEO database using the sample attribute or parameter of the GEO IDENTIFIER.

4. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Standard Attribute** menu, select the annotation you want to add.
   c. Click the **Insert Standard Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
5. To add a custom annotation to the URL field, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Custom Attribute** menu, enter the name of the sample attribute you want.
   c. Click the **Insert Custom Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
6. Click the **OK** button.

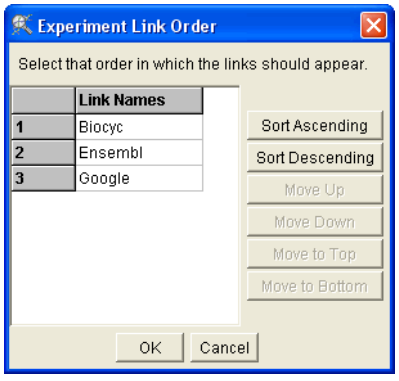
   The Import Genome: Web Links window opens.
7. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 4-22.
   - If you are done adding Web links, go to "Saving New Genomes" on page 4-27.

Setting the Experiment Link Order

To set the experiment link order:

1. In the **Import Genome: Web Links** window, click the **Set Link Order** button in the Experiment Links table.

   The Experiment Link Order window opens.



   This window lets you set the order in which genes should appear in the Condition Inspector.

2. Do the following:
   - To sort the links in ascending order, click the **Sort Ascending** button.
   - To sort the links in descending order, click the **Sort Descending** button.
   - To move the selected link up one row, click the **Move Up** button.
   - To move the selected link down one row, click the **Move Down** button.
   - To move the selected link to the top of the column, click the **Move to Top** button.
   - To move the selected link to the bottom of the column, click the **Move to Bottom** button.
3. Click the **OK** button.

   The Import Genome: Web Links window opens.
4. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 4-22.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Removing Experiment Links

To remove an experiment link:

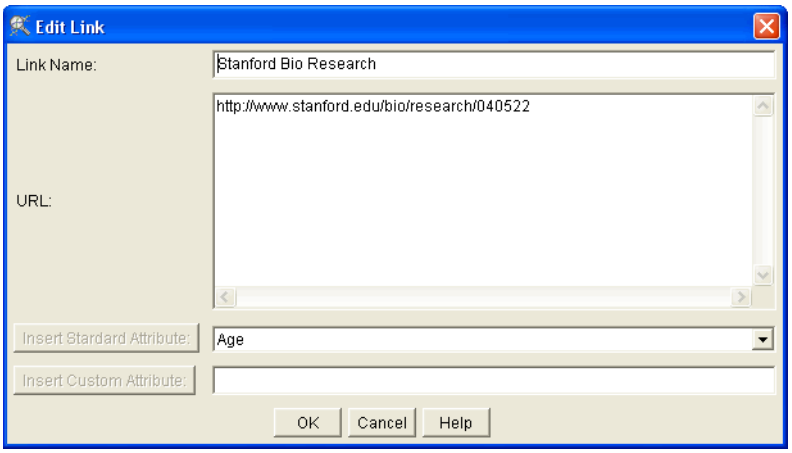
1. In the **Import Genome: Web Links** window, select a link to remove in the Experiment Links table.
2. Click the **Remove Link** button.
3. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 4-22.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Editing Experiment Links

To add an experiment link:

1. In the **Import Genome: Web Links** window, select a link to edit in the Experiment Link table.
2. Click the **Edit** button.

   The Edit Link window opens.



3. In the **Link Name** field, edit the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Condition Inspector.
4. In the **URL** field, edit the URL you want for a particular web database.
5. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Standard Attribute** menu, select the annotation you want to add.
   c. Click the **Insert Standard Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
6. To add a custom annotation to the URL field, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Custom Attribute** menu, enter the name of the sample attribute you want.
   c. Click the **Insert Custom Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
7. Click the **OK** button.

   The Import Genome: Web Links window opens.
8. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 4-22.
   - If you are done adding web links, go to "Saving New Genomes" on page 4-27.

Saving New Genomes

This section explains how to save your new genomes after you have imported them into GeneSpring.

Workflow for Saving New Genomes

To save new genomes, you perform one or more of the following steps:

1. Name the genome and save it in a GeneSpring folder.
2. Update the genome annotations from a public database.
3. Make gene lists from properties.
4. Build an ontology based on the Gene Ontology (GO SLIMS) classification system.
5. Import Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.
6. Build a homology table.
7. Add supplemental information from the Silicon Genetics web site.

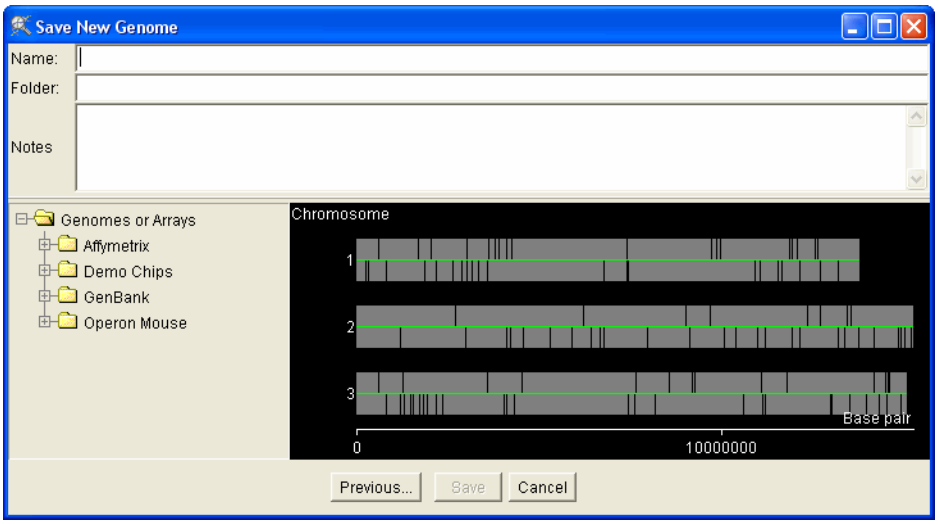
The Import New Genome wizard will guide you through each step of the process.

Naming New Genomes

To name the new genome:

1. In the **Import Genome: Web Links** window, click the **Next** button.

   The Save New Genome window opens.



   If an image of the genome is available, it displays in the image viewer. Otherwise, the message, “Image not available,” appears.

   The list of pre-made genomes is organized into folders and files. The folders list the name of the manufacturer and the folders contain the genomes from that manufacturer.

2. In the **Name** field, enter a name for the genome.

   If this genome was downloaded from Silicon Genetics, then the name field will display the default name for the genome.

3. Single-click the **Genomes and Arrays** folder.

4. Navigate to the genome folder you want.

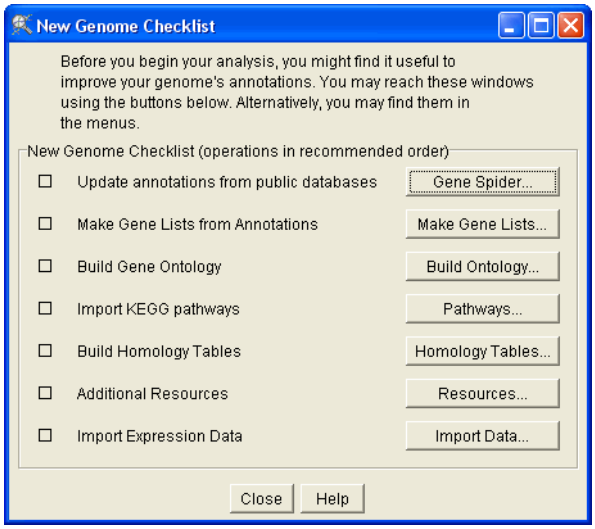
5. Single-click the folder.

   The folder name appears in the Folder field.

6. In the **Notes** field, add any information you want about the new genome.

7. Click the **Save** button.

   The genome is saved and the New Genome Checklist window opens.


New Genome Checklist
Before you begin your analysis, you might find it useful to improve your genome's annotations. You may reach these windows using the buttons below. Alternatively, you may find them in the menus.
New Genome Checklist (operations in recommended order)
Update annotations from public databases
Gene Spider...
Make Gene Lists from Annotations
Make Gene Lists...
Build Gene Ontology
Build Ontology...
Import KEGG pathways
Pathways...
Build Homology Tables
Homology Tables...
Additional Resources
Resources...
Import Expression Data
Import Data...
Close
Help

   This check list provides access to tools for improving your genome annotations before proceeding with data analysis.

8. After you have saved your genome, you can perform one or more of the following tasks:

   - To update your annotations with GeneSpider, go to "Updating Annotations with GeneSpider" on page 4-29.
   - To make gene lists from properties, go to "Making Gene Lists from Annotations" on page 4-30.
   - To build a gene ontology, go to "Building Ontologies" on page 4-31.
   - To import a KEGG pathway, go to "Importing KEGG Pathways" on page 4-32.
   - To build a homology table, go to "Building Homology Tables" on page 4-34.
   - To update supplemental information, go to "Obtaining Supplemental Information" on page 4-36
   - To import a data file, go to "Loading Experiments" on page 5-3.

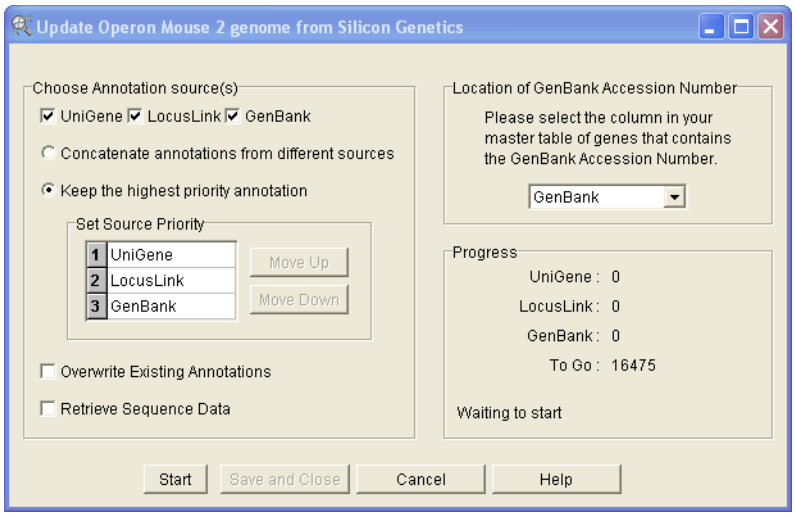
Updating Annotations with GeneSpider

The Gene Spider function retrieves annotations from the NCBI web site or the Silicon Genetics mirror. The Silicon Genetics mirror server downloads the complete databases for GenBank, RefSeq, LocusLink, and UniGene from NCBI. It returns the same annotations as the GeneSpiders that access GenBank, LocusLink, and UniGene, depending on the annotation sources chosen by the user in the Update genome from Silicon Genetics window. In addition, the Silicon Genetics GeneSpider fills in GO biological process, GO molecular function, GO cellular component, and RefSeq identifier fields from the LocusLink database and the UniGene cluster ID from the UniGene database.

To update annotations with GeneSpider:

1. In the **New Genome Checklist** window, click the **GeneSpider** button.

   The Update Genome from Silicon Genetics window opens.



   The date when GeneSpider was last run for this genome is indicated at the top of the window. If no date is shown, the GeneSpider has not been run before.

2. In the **Choose Annotation Source** box, select the annotation source and source priority options you want:

   a. Select the boxes next to the annotation sources from which to retrieve data.

   b. Select **Concatenate annotations from different sources** to retrieve annotations from all sources as a semicolon-delimited list. Exact duplicates are not retrieved. The order is fixed: GenBank, then LocusLink, and finally, UniGene.

   c. Select **Keep the highest priority annotation** to retrieve only the annotation from the highest priority source available for each gene. Use the **Move Up** and **Move Down** buttons to reorder the priority.

   d. To update information in places where data already exists, select the **Overwrite Existing Annotations** check box.

If you leave this box unchecked, GeneSpring adds new information only to blank fields. When you update annotations, GeneSpring creates a back-up file of the pre-update master gene table.

e. To update sequence data from Silicon Genetics or GenBank, select the **Retrieve Sequence Data** option.

3. In the **Location of GenBank Accession Number** box, select the column in your Master Table of Genes that contains GenBank accession numbers.
4. Click the **Start** button.

   Depending on the number of genes in the genome, this process can take a number of hours to complete. Therefore, it is recommended to run this function over night.

   While the GeneSpider runs, there are a number of informational fields visible.

   - **Status**—this is the level of completion the GeneSpider has reached.
   - **GenBank, UniGene, LocusLink**—Number of genes whose annotations have been retrieved from the given database.
   - **Processed**—Number of genes in the genome that the GeneSpider has finished querying the database.
   - **Found**—number of processed genes where the GeneSpider has found a useful record in the database.
   - **Enhanced**—Number of genes where information has been found and added to the Master Table of Genes.
   - **To Go**—Number of genes in the genome that have not been processed for the current database.

   The Master Table of Genes is not updated until you click **Save and Close**. This button is inactive while the GeneSpider is running.
5. You must wait until the GeneSpider is finished, or click the **Stop** button, before clicking **Save and Close**.

   After the process successfully completes, the New Genome Checklist window opens with this option selected.
6. If you are done working in the New Genome Checklist window, click the **Close** button.

Making Gene Lists from Annotations

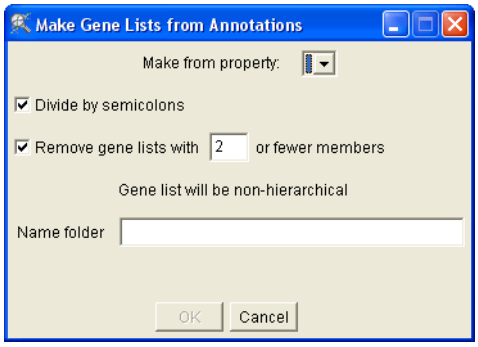
During data analysis, you work with interesting collections of genes known as gene lists. These gene lists are stored in the Gene Lists folder. By default, GeneSpring makes and displays an “all genes” list containing all genes in the genome. You can also create your own gene lists based on the annotation values for each gene.

When you run this tool, GeneSpring creates a gene list for each of the values in the annotation column. This will allow you to group those genes that have similar values for a particular annotation. This tool is particularly useful when you need to group genes with similar functions or classifications.

To make gene lists from annotations:

1. In the **New Genome Checklist** window, click the **Make Gene Lists** button.

   The Make Gene List from Annotations window opens.



2. In the **Make from Property** menu, choose the annotation column you want to use for generating the gene list.

   After making a selection, the **Name Folder** field is automatically filled in for you.

3. Select the **Divide by Semicolons** check box.

4. Click the **OK** button.

   The folder that will be created will contain all the gene lists in a non-hierarchical manner. The New Genome Checklist window opens with this option checked off.

5. If you are done working in the New Genome Checklist window, click the **Close** button.

Building Ontologies

GeneSpring 7 uses the Gene Ontology (GO SLIMS) classification system to build ontologies. The GO Consortium has created a controlled vocabulary that can be applied to all organisms (even as knowledge of gene and protein roles in cells is changing).

GO SLIMS provides a high-level view of gene ontology classification by limiting the number of GO classifications to about a hundred. GO SLIMS creates fewer gene lists than the original ontology classification, but captures more genes, which results in better ontologies.This will allow you get a good sense of a gene's GO classification, without the sheer complexity of the full gene ontology classifications (which contains about 17,000 different classifications).

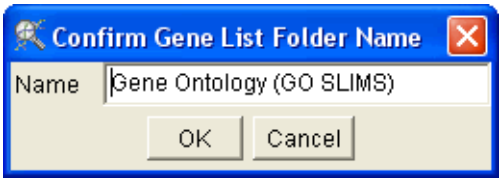
The Build Ontology tool creates a gene list for each of the ontological classifications. This tool will group all the genes with the same GO classification into the same gene list. After the gene lists have been created, the Similar Lists function in the Gene List Inspector can be used to find overlap between your gene list of choice and the GO gene lists. This information can be used to determine whether your gene lists contains an over-representation of genes from one or more GO classifications—which may lead to conclusions about the genes in the gene list.

Using gene lists to represent the GO classifications can also be used to determine GO classifications for a single gene in the Gene Inspector. Go to "Using the Gene Inspector" on page 7-16 for more information.

To build a new ontology:

1. In the **New Genome Checklist** window, click the **Build Ontology** button.

   The Confirm Gene List Folder Name window opens.



2. Enter the name of the folder in which you want to store the GO gene list.
3. Click the **OK** button.

   If successful, GeneSpring displays the message, "Successfully built ontology. The New Genome Checklist window opens with this option checked off.
4. If you are done working in the New Genome Checklist window, click the **Close** button.

Importing KEGG Pathways

The Kyoto Encyclopedia of Genes and Genomes is a repository for gene and protein information that is maintained in Japan (http://www.genome.ad.jp/kegg/). Part of KEGG is a repository of Pathway diagrams that depict many different types of biological pathways, both Metabolic and Regulatory pathways.

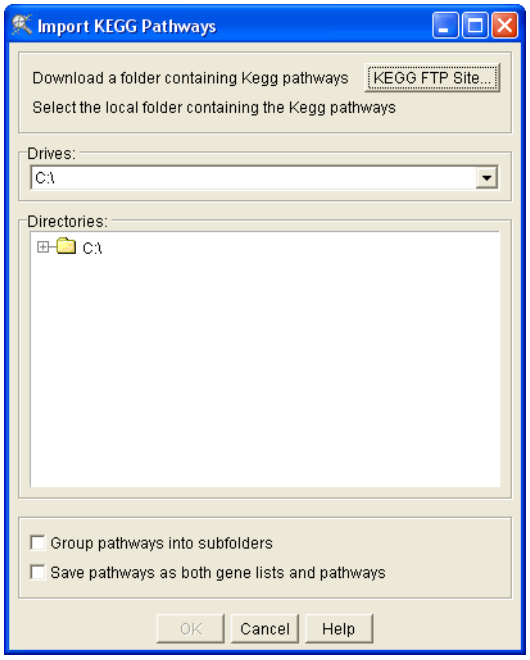
These pathway diagrams can be imported into GeneSpring and the expression values for each gene can be superimposed on those diagrams, to provide you with a direct insight in the behavior of the gene in the expression experiment and their place in the pathways. You can utilize the KEGG pathways by downloading the pathway diagrams from the KEGG FTP site and loading them into GeneSpring.

Note: KEGG requires a license agreement for commercial entities and is only freely available for academic users. Please do not download KEGG unless you have the appropriate license and Silicon Genetics can not be held responsible for any illegal use of the KEGG pathways.

To import KEGG pathways:

1. In the **New Genome Checklist** window, click the **Pathways** button.

   The Import KEGG Pathways window opens.



2. To import a pathway from the KEGG FTP site, do the following:

   a. Click the **FTP KEGG** button.

      This button links you to ftp://ftp.genome.ad.jp/pub/kegg/pathways.

   b. Select the organism-specific folder (recommended) or the folder called map, which contains generic metabolic pathways.

      Organism-specific folders are named by the first letter of the genus and first two letters of the species (for example, 'hsa' stands for Homo sapiens).

   c. To download a folder from the FTP site, either right-click on the folder and select **Copy to Folder,** or drag the folder to your hard drive.

3. To import a pathway on your system, do the following:

   a. In the **Drive** menu, select the drive you want.

   b. In the **Directories Navigator**, select the directory you want.

   c. Select the pathway folder you downloaded from the KEGG site in step 2c.

4. To group the pathways into subfolders, select the **Group the pathways into subfolders** checkbox.

   This option can be useful if there are many pathways in the folder, and is recommended if there are more than 100 pathways in the KEGG pathway folder you wish to import. You may also choose to save pathways as both Pathway images and gene lists.

Note: The KEGG map folder of pathways will only contain EC numbers to allow mapping genes onto the associated pathway images. All other KEGG pathways contain additional information allowing GeneSpring to map genes using their Common Name, GenBank accession number, and if a match has not been found, using EC numbers.

5. To save pathways as gene lists and pathways, select the **Save pathways as both gene lists and pathways** check box.

   This option can be useful when you want to represent the pathways as gene list. Each gene list will have the same name as the pathway it was created from, and will contain all the genes that are represented in that pathway.

   Once genes are also represented as a pathway, you can quickly determine if a gene is a member of a pathway by examining the "Lists Containing gene" feature in the Gene Inspector. Alternatively, you can determine if a gene list you created contains an over-representation of genes belonging to specific pathways. Go to "Using the Gene List Inspector" on page 7-33 for more information.

6. Click the **OK** button.

   The New Genome Checklist window opens with this option checked off.

7. If you are done working in the New Genome Checklist window, click the **Close** button.

Building Homology Tables

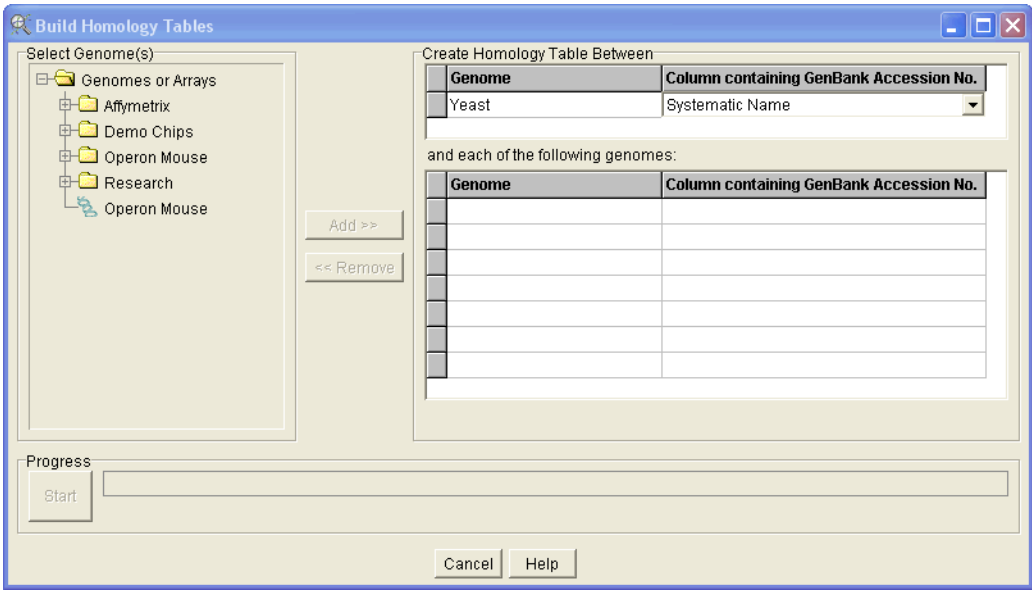
The Homology tool automates the process of building homology tables for certain organisms. Currently, a limited list of organisms is available. Using this tool, homologies can be made between any pair of organisms that are included in both HomoloGene and UniGene. Within-genome homologies are based solely on UniGene Cluster ID or LocusLink Locus ID.

This section provides a brief overview of how to build a new homology table. Go to "Working with Homology Tables" on page 12-23 in for more information.

To build an homology table:

1. In the **New Genome Checklist** window, click the **Homology Tables** button.

   The Build Homology Tables window opens.



2. From the menu in the **Column Containing GenBank Accession No.**, select the appropriate column that contains the GenBank (or EMBL) accession number for the selected genome.
3. In the **Navigator**, select the genome you want and click **Add**.

   The genome is added to the lower table on the right side of the window.
4. From the menu next to the newly added genome, select the name of the column containing the genome's GenBank Accession Number.
5. Repeat steps 2 through 4 for each genome you want to add.
6. Click the **Start** button.

   This process can take several hours to complete. If the initially selected column does not have GenBank Accession Numbers, an error message appears. If a selected genome is not on Homologene, you will receive an error message after the Homology tool has finished running.
7. When prompted, specify whether or not to save the UniGene Cluster IDs.

   When you choose to select the UniGene ID's, they will be saved in a annotation column called "Unigene". Any data that was previously in the Unigene column will be overwritten.

   The resulting homology tables are saved in both genomes. The New Genome Checklist window opens with this option checked off.
8. If you are done working in the New Genome Checklist window, click the **Close** button.

Obtaining Supplemental Information

You can download supplemental information for use in GeneSpring from the Silicon Genetics web site. This supplemental information consists of many different types of information, like new pathway diagrams, chromosome maps, or sequence information that can be used in the Find Regulatory Sequence search function. Go to "Viewing Regulatory Sequence Search Results" on page 14-6 for more information.

To obtain supplemental information:

1. In the **New Genome Checklist** window, click the **Resource** button.

   The Silicon Genetics web site opens.

2. Search the web site to locate supplemental information to download. For example:

   - A zip installer containing mouse chromosome maps
   - A zip installer containing human chromosome maps
   - A zip installer containing GenMAPP pathways

3. After you locate and download the information you want, the New Genome Checklist window opens with this option checked off.

4. If you are done working in the New Genome Checklist window, click the **Close** button.

5

Working With Experiments

This chapter explains how to set up an experiment in GeneSpring. It covers the following topics:

- Before You Begin
- Loading Experiments
- Creating New Experiments
- Copying and Pasting Experiments
- Using the Default Normalizations
- Setting Up Experiment Parameters
- Working with Samples
- Working with Sample Attributes
- Setting Up Experiment Interpretations
- Using Cross-Gene Error Models

Before You Begin

Before working with experiments in GeneSpring, please review the information in this section.

File Formats

GeneSpring can load data from nearly any expression analysis technology, provided the data is formatted as tab-delimited text. GeneSpring recognizes various file formats from commercially-available products. These products include the following:

- Affymetrix Pivot
- Affymetrix Metrix
- Affymetrix MAS 5.0
- Agilent Feature Extraction
- Amersham Codelink
- Amersham Expression Report
- Amersham GeneSpring Report
- Axon GenePix Pro 2 & 3
- BioDiscovery ImaGene 4.0

- Clontech Two Color
- Clontech One Color
- Harvard (Wong Lab) dChip
- Incyte Internet
- Incyte GEM Tools 2.4
- Clontech AtlasImage 2.0

If GeneSpring is unfamiliar with your file format, you can define a custom format to specify the type of data in each column. These specifications can be added to the list of known file types so that you can load subsequent experiments in batches.

Make sure you use the raw, tab-delimited files just as they come out of the scanner. GeneSpring uses the information in the column headers. If you have cut out header information, use your original tab-delimited data files.

Memory Requirements for Loading Experiments

In GeneSpring 6.0 and later, experiments are loaded into the disk cache as well as into system memory (RAM). This requires some additional time when an experiment is first created. Once the experiment is loaded, it can then be reloaded in a fraction of the time. This change was made to accommodate the loading and creation of very large experiments, especially for systems with limited memory.

If you want to free some hard disk space or think that your cached data folder may be corrupted, delete the cache folder (`genespring/data/cache`, where "GeneSpring" is the GeneSpring home directory on your machine). This forces GeneSpring to recreate the experimental data, which may solve the problem.

Using Data Preprocessors

In GeneSpring 7, you can now write data preprocessors using the GeneSpring plugin API. A plugin lets you pre-process data files before creating samples for an experiment. By calculating gene intensity from the original probe-level measurements, you no longer need to re-import samples that have been pre-processed by other probe-level analysis packages.

To use data preprocessors in GeneSpring, you (or a developer) must write the plugin and then install it in GeneSpring. Once the data preprocessor is installed, it will automatically show up in GeneSpring when you attempt to load files that are readable in a format that the preprocessor recognizes.

For information on installing data preprocessors, go to "Working with Data Preprocessor Plugins" on page 16-63. If you are a developer and would like more information on the data preprocessor API, contact Contact Silicon Genetics Technical Support at 1-866-SIG-SOFT.

Loading Experiments

This section explains how to load experiments in GeneSpring. It covers the following topics:

- Loading Expression Data from Files
- Selecting Signal and Control Files
- Selecting Samples with Multiple Files
- Managing Data Files with Extra Genes
- Importing Sample Attributes
- Saving New Experiments

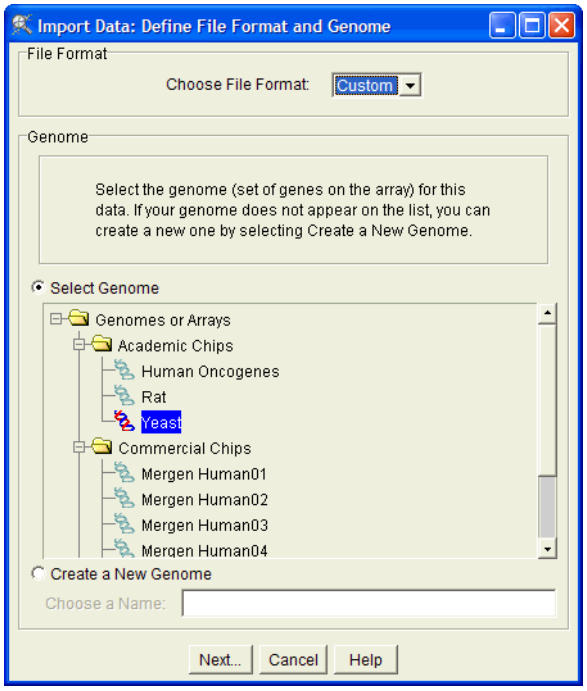
Loading Expression Data from Files

To load expression data from a file:

1. Do one of the following:
   - Select **File** > **Import Data**, select the file you want, and then select **Open**.
   - Press **Ctrl+O**, select the file you want, and then select **Open**.
   - (Windows and UNIX) Drag and drop the files you want from your desktop (or other location) into the main GeneSpring window.

     Note: You can select multiple files using the drag-and-drop option.

     The Import Data: Define File Format and Genome window opens.



2. In the **Choose File Format** box, select the correct file format from the menu (if it is not already selected).

   GeneSpring will examine the file and attempt to guess the format. The Choose File Format menu will default to the most appropriate format. It will not list all possible formats, but only those formats that are compatible with the imported file. The **Custom format** option is always available and should be used if the default choice is not appropriate, or if the format is not recognized by GeneSpring.

3. Do one of the following:

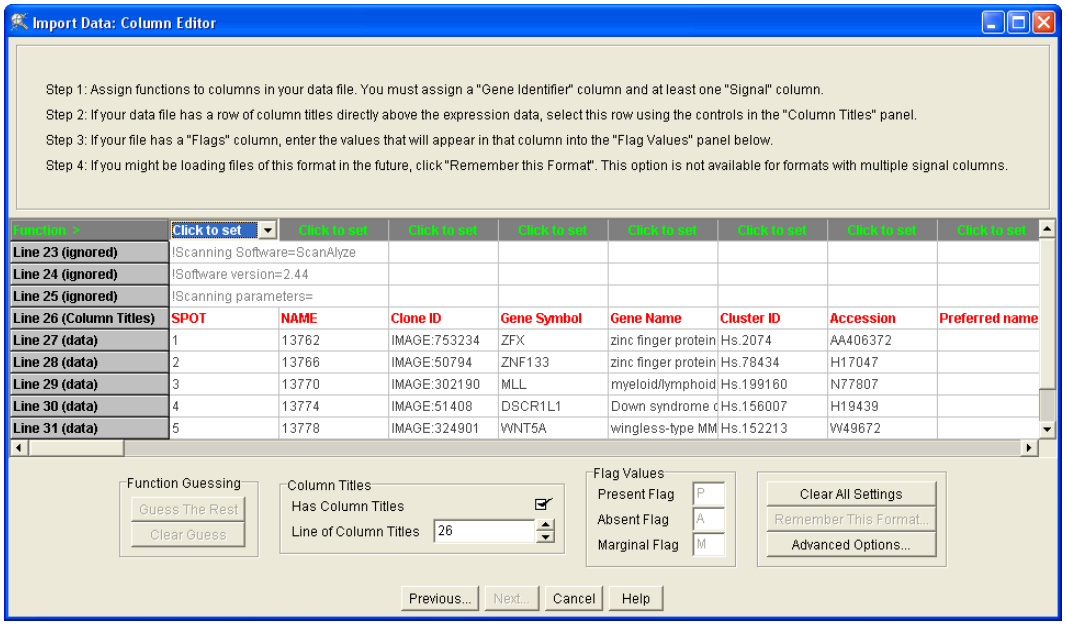
   - To choose a genome that has previously been loaded, select the **Select Genome** option, and then select the genome from the folder or file you want.
   - If a genome representing the array has not previously been created, you can create one based on the data in the file you are loading. Select the **Create a New Genome** option and enter the name you want in the **Choose a Name** field.

     Note: The disadvantage of this method is the fact that no gene annotation, other than the annotation in the file, will be available during the analysis. You can, however, add an annotation at a later time. Go to "Importing Genes and Annotations" on page 6-46 for information.

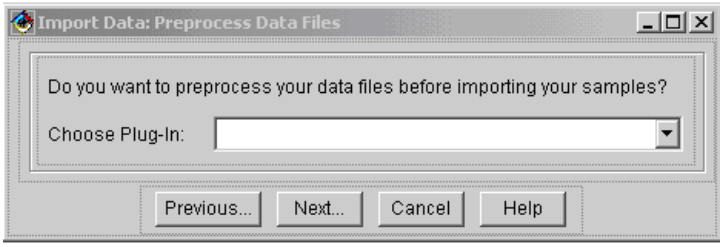
4. Click the **Next** button.

5. Do one of the following:

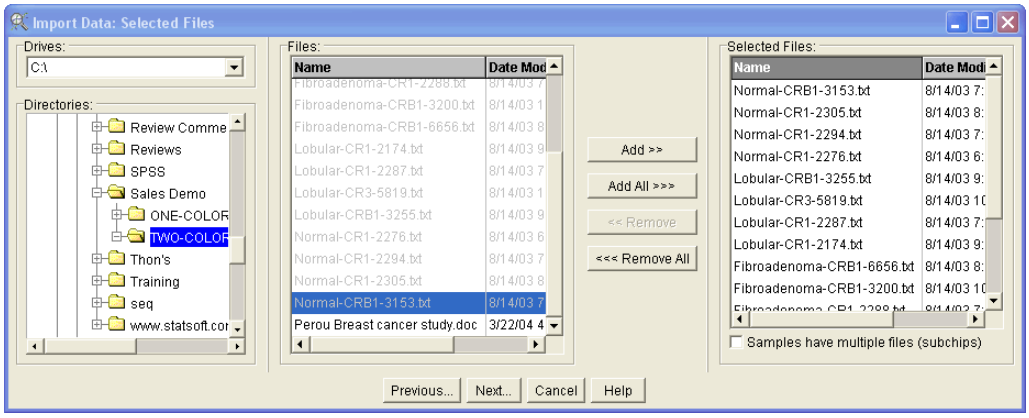
   - If your data is in a custom format, the Import Data: Column Editor window opens. You must set up columns before continuing. Go to "Using the Column Editor" on page 5-5 to continue.


Import Data: Column Editor
Step 1: Assign functions to columns in your data file. You must assign a "Gene Identifier" column and at least one "Signal" column.
Step 2: If your data file has a row of column titles directly above the expression data, select this row using the controls in the "Column Titles" panel.
Step 3: If your file has a "Flags" column, enter the values that will appear in that column into the "Flag Values" panel below.
Step 4: If you might be loading files of this format in the future, click "Remember this Format". This option is not available for formats with multiple signal columns.
Function >
Click to set
Line 23 (ignored) !Scanning Software=ScanAlyze
Line 24 (ignored) !Software version=2.44
Line 25 (ignored) !Scanning parameters=
Line 26 (Column Titles) SPOT NAME Clone ID Gene Symbol Gene Name Cluster ID Accession Preferred name
Line 27 (data) 1 13762 IMAGE:753234 ZFX zinc finger protein Hs.2074 AA406372
Line 28 (data) 2 13766 IMAGE:50794 ZNF133 zinc finger protein Hs.78434 H17047
Line 29 (data) 3 13770 IMAGE:302190 MLL myeloid/lymphoid Hs.199160 N77807
Line 30 (data) 4 13774 IMAGE:51408 DSCR1L1 Down syndrome Hs.156007 H19439
Line 31 (data) 5 13778 IMAGE:324901 WNT5A wingless-type MM Hs.152213 W49672
Function Guessing
Guess The Rest
Clear Guess
Column Titles
Has Column Titles
Line of Column Titles 26
Flag Values
Present Flag P
Absent Flag A
Marginal Flag M
Clear All Settings
Remember This Format...
Advanced Options...
Previous...
Next...
Cancel
Help

- If a preprocessor plugin has been installed that recognizes the data you are loading, the Import Data: Preprocess Files window opens. This window lets you choose whether to run a data file preprocessor plugin, and if so, which one. Go to "Preprocessing Data Files" on page 5-13 to continue.



- If your data is in a known format, the Import Data: Selected Files window opens. This window lets you select more files of the same type to add to your experiment. Go to "Importing Selected Files" on page 5-14 to continue.



Using the Column Editor

If GeneSpring does not recognize your file format, use the Column Editor to assign headings and functions to each column in your data file. This section explains how to use the Column Editor and covers the following topics:

- Column Editor Window
- Default Column Assignments of Known Products
- Opening the Column Editor
- Setting Up Column Titles
- Setting Up Columns

Column Editor Window

The Column Editor window (Figure 5-1) lets you assign functions to the data file, column titles to expression data, and values to flags (if you have them).

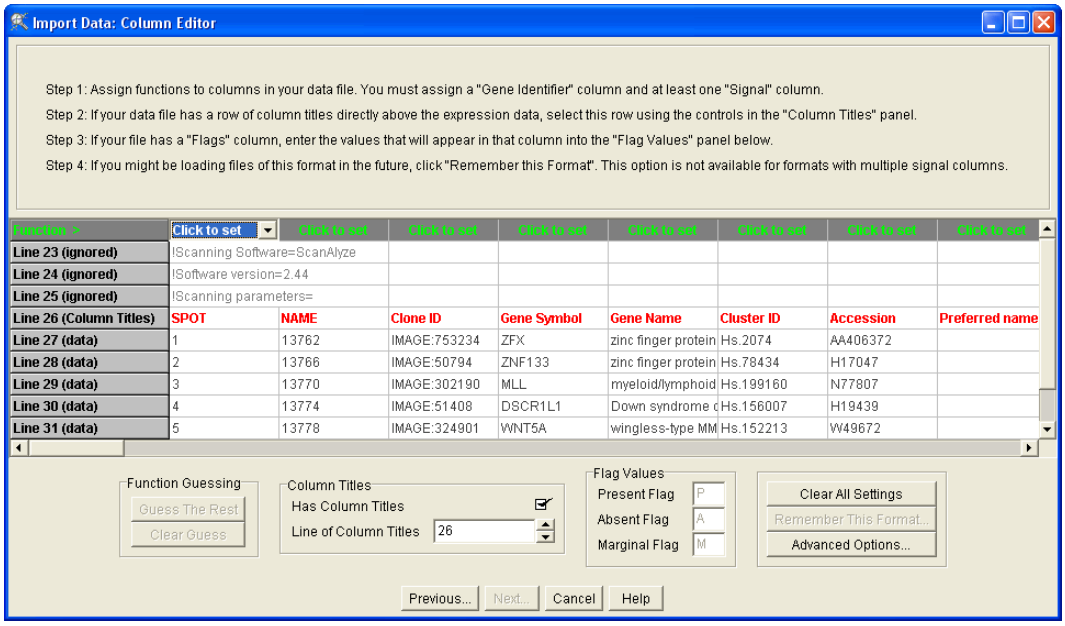

Import Data: Column Editor
Step 1: Assign functions to columns in your data file. You must assign a "Gene Identifier" column and at least one "Signal" column.
Step 2: If your data file has a row of column titles directly above the expression data, select this row using the controls in the "Column Titles" panel.
Step 3: If your file has a "Flags" column, enter the values that will appear in that column into the "Flag Values" panel below.
Step 4: If you might be loading files of this format in the future, click "Remember this Format". This option is not available for formats with multiple signal columns.
Function >
Click to set
Line 23 (ignored) !Scanning Software=ScanAlyze
Line 24 (ignored) !Software version=2.44
Line 25 (ignored) !Scanning parameters=
Line 26 (Column Titles) SPOT NAME Clone ID Gene Symbol Gene Name Cluster ID Accession Preferred name
Line 27 (data) 1 13762 IMAGE:753234 ZFX zinc finger protein Hs.2074 AA406372
Line 28 (data) 2 13766 IMAGE:50794 ZNF133 zinc finger protein Hs.78434 H17047
Line 29 (data) 3 13770 IMAGE:302190 MLL myeloid/lymphoid Hs.199160 N77807
Line 30 (data) 4 13774 IMAGE:51408 DSCR1L1 Down syndrome c Hs.156007 H19439
Line 31 (data) 5 13778 IMAGE:324901 WNT5A wingless-type MM Hs.152213 W49672
Function Guessing
Guess The Rest
Clear Guess
Column Titles
Has Column Titles
Line of Column Titles 26
Flag Values
Present Flag P
Absent Flag A
Marginal Flag M
Clear All Settings
Remember This Format...
Advanced Options...
Previous...
Next...
Cancel
Help

Figure 5-1 Import Data: Column Editor Window

Default Column Assignments of Known Products

GeneSpring recognizes various commercially-available products and places them, as described, in the following lists.

Affymetrix

Pivot Table:

- **Column 1—**interpreted as Gene Name
- **Average Difference or Signal**—interpreted as Signal
- **Detection or Abs Call**—interpreted as Flags

Metrix:

- **Gene Name or Probe Set or Probe Set Name—**interpreted as Gene Name
- **Signal or Average Difference—**interpreted as Signal
- **Detection or Abs Call—**interpreted as Flags
- **P, M, A—**interpreted as Flag Designators
- **Region—**interpreted as Experiment Name

d-Chip

- **Probe Set**—interpreted as Gene Name
- **Column to left of column that ends in "call"**—interpreted as Signal
- **Description**—interpreted as Description
- **Accession**—interpreted as GenBank ID
- **Column to the right of "call" that is to the right of a Signal column**—interpreted as Flags
- **P, M, A**—interpreted as Flag Designators

Agilent

- **ProbeName**—interpreted as Gene Name
- **rBGSubSignal**—interpreted as Signal
- **gBGSubSignal**—interpreted as Control
- **Description or GeneName**—interpreted as Description
- **GenBank**—interpreted as GenBank ID

Amersham

- **GeneID**—interpreted as Gene Name
- **Signal Mean**—interpreted as Signal
- **Background Mean**—interpreted as Signal Background
- **Flag**—interpreted as Flags
- **0=P, 2=A, 3=M**—interpreted as Flag Designators

Axon

GenePix Pro 2 & 3:

The Ratio Formulation entry is used to determine which channel is Signal and which is Control.

- **ID—**interpreted as Gene Name
- **F635 Median or F532 Median—**interpreted as Signal
- **B635 Median or B532 Median—**interpreted as Signal Background
- **F635 Median or F532 Median—**interpreted as Control Channel
- **B635 Median or B532 Median—**interpreted as Control Channel Background
- **Name**—interpreted as Description
- **Flags**—interpreted as Flag

BioDiscovery

Imagene 4:

- **Gene ID—**interpreted as Gene Name
- **Signal Median—**interpreted as Signal
- **Background Median—**interpreted as Signal Background
- **Signal Median—**interpreted as Control Channel
- **Background Median—**interpreted as Control Channel Background
- **Flag**—interpreted as Flag

Incyte

GEMTools 2.4:

- **CloneID—**interpreted as Gene Name
- **P2 BalancedSignal or P2 Balanced—**interpreted as Signal
- **P1Signal or P1—**interpreted as Control Channel
- **Gene Name—**interpreted as Description
- **AccessionNum or Accession—**interpreted as GenBankID

Internet Download:

- **CloneID—**interpreted as Gene Name
- **Varies (format of PS# Cy5, determined in header)—**interpreted as Signal
- **Varies (format of PS# Cy3, determined in header)—**interpreted as Control Channel
- **Gene name—**interpreted as Description
- **PS# Absent/Present (where # is the sample name)—**interpreted as Flags
- **P, A—**interpreted as Flag Designators
- **GEM ID—**interpreted as Custom1
- **Gene ID—**interpreted as Custom2

Packard Biochip ScanArray/QuantArray

Check the file header to determine which channel is Signal and which is Control.

- **Name—**interpreted as Gene Name
- **ch1 Intensity or ch2 Intensity—**interpreted as Signal
- **ch1 Background or ch2 Background—**interpreted as Signal Background
- **ch1 Intensity or ch2 Intensity—**interpreted as Control Channel
- **ch1 Background or ch2 Background—**interpreted as Control Channel Background

Clontech Atlas Image 2-Color

- **Gene Code—**interpreted as Gene Name
- **Intensity_2—**interpreted as Signal
- **Background_2—**interpreted as Signal Background
- **Intensity_1—**interpreted as Control Channel
- **Background_1—**interpreted as Control Channel Background
- **Protein/gene—**interpreted as Description
- **Column 11—**interpreted as GenBankID

Clontech Atlas Image 1-Color

- **Gene Code—**interpreted as Gene Name
- **Intensity_2—**interpreted as Signal
- **Background_2—**interpreted as Signal Background
- **Protein/gene—**interpreted as Description
- **Column 11—**interpreted as GenBankID

Opening the Column Editor

When you attempt to load expression data that is either not recognized by GeneSpring or is in a custom format, the Column Editor window opens. You must set up columns before continuing with the import process.

Setting Up Column Titles

To set up column titles:

1. When you first load a file, GeneSpring analyzes it to determine which row contains the column titles. If the row chosen is incorrect, use the **Line Containing Column Titles** field to adjust the number of rows.
2. If there are no column titles in your data file, clear the box marked **Has column titles**.
3. To assign functions to each data column, choose a function from the menu.

Setting Up Columns

To set up columns:

1. Designate at least one Gene Identifier column and one Signal (raw data) column.

   Refer to Table 5-1 for a description of the available column assignments.

Table 5-1 Column Editor Column Assignments

Name	Required?	# Allowed	Description
Unused	Optional	Any	These columns are not used to load any data into GeneSpring. They can also be used to filter data via the Filter Genes window. See "Filtering on Data Files" on page 9-23 for details.
Gene Identifier	Required	One	This column is used for the Systematic Name. Gene identifiers must be unique to the genes in this genome. Duplicate genes are treated as replicates. The mean of the expression value for all on-chip replicates is calculated and used as the expression value for the gene. It is recommended that the Gene Identifier in the raw data files be the gene's Systematic Name.
Signal	Required	One or more	This column contains the (raw) expression values for each of the genes. You must have at least one Signal column.
Signal Background	Optional	Any	This column contains the background signal. You can have as many Signal Background columns as you have Signal columns. If you are using Signal Background, you must have a Signal Background for each Signal column. During data import, GeneSpring automatically performs a background correction of the signal value by subtracting the background signal from the signal value.
Signal Precision	Optional	Any	Used only when the scanner software used for your experiment produces an estimate of the precision of the value in the signal column. This information is merged with other information as part of the GeneSpring Cross-Gene Error Model. These numbers are the standard deviation of the measured signal around the true expression level (signal) for that sample as expressed by the scanner software. See "Using Cross-Gene Error Models" on page 5-73 for details.
Control Channel	Optional	Any	This column contains the expression values for the second channel in two-color expression technologies. In GeneSpring, the second channel is called the control channel, since in most cases the second channel is used as the control sample or hybridization. If you have control channels (for example, a two-color experiment), you must have the same number of control channel columns as signal columns.
Control Channel Background	Optional	Any	This column contains the background signal for the control channel. If you are using control channel backgrounds, the number of columns must be the same as the number of Control Channel columns.

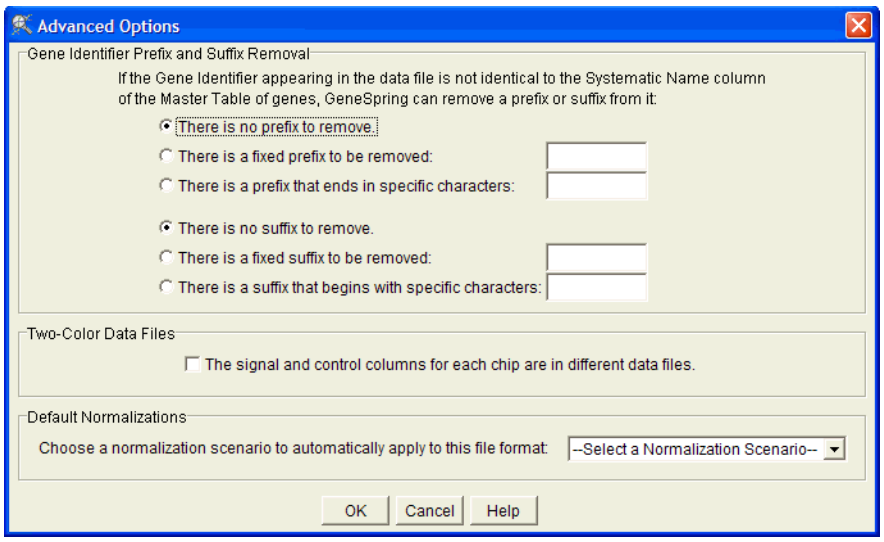
Table 5-1 Column Editor Column Assignments (Continued)

Name	Required?	# Allowed	Description
Description	Optional	One	This column contains a description of the gene, if known. This information is included in the new master table of genes if a new genome is created during data import. It is accessible with the Find Gene command and the Gene Inspector. This field applies only to new genomes created through the Column Editor.
GenBankID	Optional	One	This column contains the GenBank accession number for the gene, if known. GenBank accession numbers let you update information about the gene directly from GenBank. See "Updating Annotations with GeneSpider" on page 11-4 for more information. This field is included in the new Master Table of Genes and applies only to new genomes created through the Column Editor.
Common Name	Optional	One	This column contains a gene name or some other commonly used name to refer to a gene. This information is included in the new Master Table of Genes when a new genome is created during data import.
Flags	Optional	Any	This column specifies the letter or number indicating Present, Absent, and Marginal calls. You can have as many Flag columns as you have Signal columns.
Region	Optional	One	If your sample uses multiple arrays, or sections of arrays that must be normalized separately, this column tells GeneSpring the region of the array or from which array a particular gene reading came. When the region column is used during import, the normalization algorithms uses this information when normalizing the data.

2. To enable GeneSpring to label the remaining columns:
   - Click the **Guess the Rest** button.

     Note: You must define two complete sets of columns for the **Guess the Rest** function to guess the pattern.
   - If the guesses are incorrect, click the **Clear Guesses** button.
3. Click the **Advanced Options** button if any of the following statements are true:
   - The gene identifiers in your experiment files have a prefix or suffix that must be stripped.
   - Your signal and control values are in separate files.
   - You want to apply a default normalization scheme to your experiment files.

From this window, you can select the appropriate options.


Advanced Options
Gene Identifier Prefix and Suffix Removal
If the Gene Identifier appearing in the data file is not identical to the Systematic Name column of the Master Table of genes, GeneSpring can remove a prefix or suffix from it:
There is no prefix to remove.
There is a fixed prefix to be removed:
There is a prefix that ends in specific characters:
There is no suffix to remove.
There is a fixed suffix to be removed:
There is a suffix that begins with specific characters:
Two-Color Data Files
The signal and control columns for each chip are in different data files.
Default Normalizations
Choose a normalization scenario to automatically apply to this file format:
--Select a Normalization Scenario--
OK
Cancel
Help

a. **To strip a gene identifier prefix or suffix**—in the Gene Identifier Prefix and Suffix Removal section, select the appropriate option and enter the characters to be stripped in the text box next to your choice.

b. **To specify that your signal and control values are in separate files**—select the box in the Two-Color Data Files section.

c. **To apply a default normalization scenario to your experiment files**—select the appropriate scenario from the menu in the Default Normalizations section. For more information on the available default normalizations, see "Using the Default Normalizations" on page 5-35.

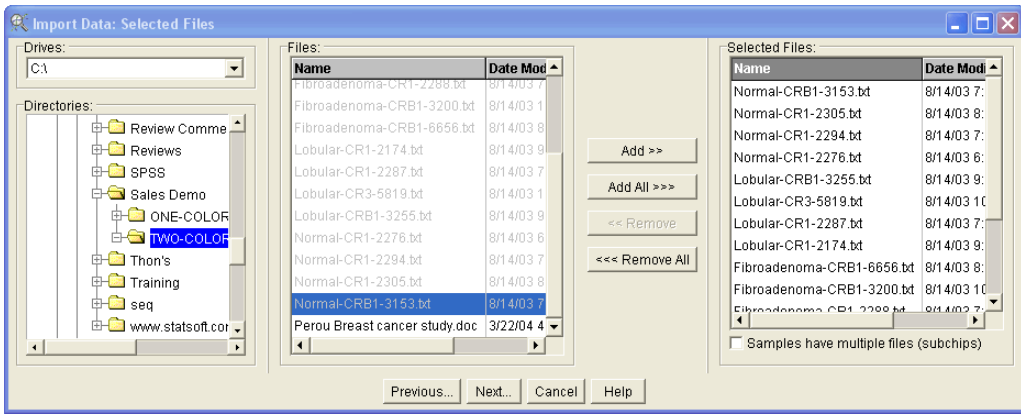
4. To save this file format setup for future use, click the **Remember This Format** button. The format is added to the cache of recognized formats so that GeneSpring recognizes it.

   Note: Formats can only be saved when there is one sample per file.

5. When prompted, enter a name for the new format.

6. Click the **OK** button.

The Import Data: Selected Files window opens.

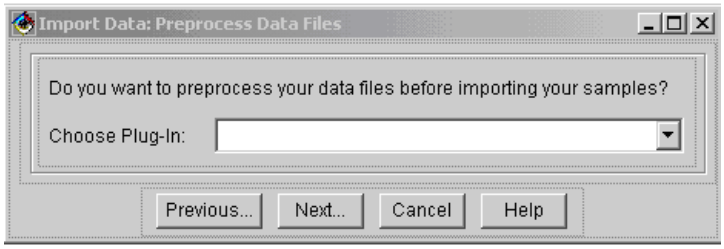


7. Go to “Importing Selected Files” on page 5-14 to continue.

Preprocessing Data Files

To preprocess your data files:

1. In the **Choose Plugin** menu, select the plugin you want to use.



2. Click the **Next** button.

   The data preprocessing starts. If the preprocessor plugin requires additional information, more windows may open and prompt you for input. For example, you might be asked to define the location of the CDF file for use in an RMA normalization pre-processor (which will be released in GeneSpring 7.1).

3. Click the **Close** button.

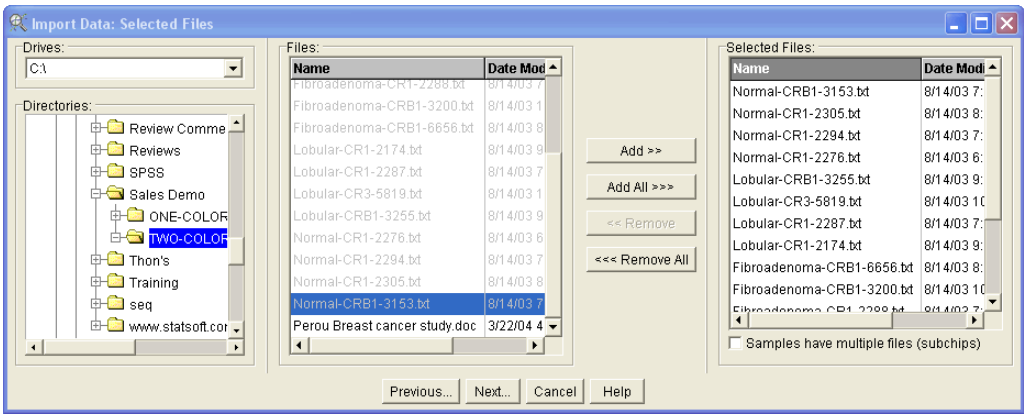
   The Import Data: Sample Attributes window opens.

4. Go to “Importing Sample Attributes” on page 5-19 to continue.

Importing Selected Files

To add files to your experiment:

1. In the Import Data: Selected Files window, select the **Drive** you want.



2. In the **Directories Navigator**, select the directory you want.
3. In the **Files Navigator**, do one of the following:
   - To sort by columns, click the column title.
   - To select one file, single-click the file.
   - To select several files at once, press Shift+click or Ctrl+click.
   - To deselect a file, press Ctrl+click.
4. Click the **Add**>> button to add the selected files, or click the **Add All**>> button to add all files.

   The selected files appear in the Selected Files column.

   Note: These files must all be in the same format. GeneSpring verifies whether the format is correct, and if it is not, it does not add the files to your experiment.
5. If you have separate subchips that you want to combine into one chip, select the **Samples have multiple files (subchips)** check box.
6. Click the **Next** button.
7. Do one of the following:
   - If a gene identifier specified during import is not unique, the Ambiguous Gene Identifiers window opens. Go to "Handling Ambiguous Gene Identifiers" on page 5-15 to continue.
   - If your signal and control files are in separate files, the Select Corresponding Files window opens. Go to "Selecting Signal and Control Files" on page 5-16 to continue.
   - If you selected the **Samples have multiple files (subchips)** check box, the Import Data: Merge Files window opens. Go to "Selecting Samples with Multiple Files" on page 5-17 to continue.

- If the genome does not contain all the genes listed in your data files, the Extend Genome window opens. Go to "Managing Data Files with Extra Genes" on page 5-18 to continue.
- If you defined recommended or required attributes, the Import Data: Sample Attributes window opens. Go to "Importing Sample Attributes" on page 5-19 to continue.

Handling Ambiguous Gene Identifiers

When the gene identifier specified during data import is not unique to a single gene in the genome, GeneSpring cannot determine the gene measurement to use. In this case, the identifier and all corresponding genes are listed in the Ambiguous Gene Identifiers window (Figure 5-2) and the measurement is not loaded.

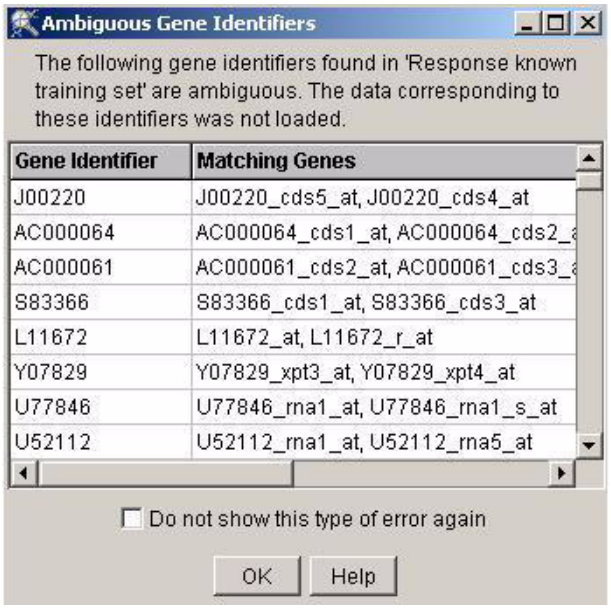


Figure 5-2 Ambiguous Gene Identifiers Window

To prevent this problem, edit the raw data file(s), assign a unique identifier to each gene in the genome (the systematic gene name in the genome is always unique), and then re-import your data. Contact Silicon Genetics Technical Support at 1-866-SIG-SOFT if you have further questions or experience difficulties.

Selecting Signal and Control Files

The Select Corresponding Files window (Figure 5-3) lets you specify a signal file and its corresponding control file. This procedure applies only to Imagene files.

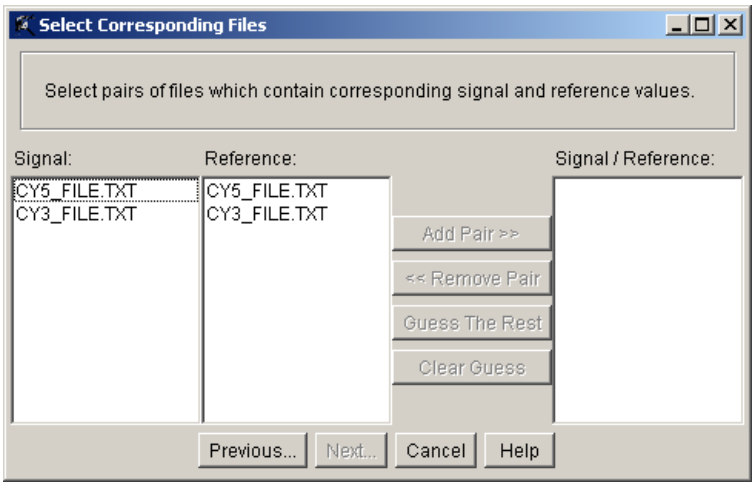


Figure 5-3 Select Corresponding Files Window

To select corresponding files:

1. In the **Signal** column, select the file name you want.
2. In the **Reference** column, select the corresponding file name.
3. Click the **Add Pair** button.

   The pair you specified appears in the Signal/Reference list.
4. To enable GeneSpring to try to select the corresponding files:
   - Click the **Guess the Rest** button.
   - If the guesses are incorrect, click the **Clear Guesses** button.
5. To remove a pair, select it in the **Signal/Reference** list and click the **Remove Pair**.
6. Click the **Next** button.
7. Go to "Importing Sample Attributes" on page 5-19.

Selecting Samples with Multiple Files

If your selected samples represent the expression values from two or more sub-chips (like the Affymetrix HG_U95v2A and HG_U95vsB), you can choose to combine these subchips into one GeneSpring sample. You use the Merge Files window (Figure 5-4) to combine the subchips from the same biological samples into one sample.

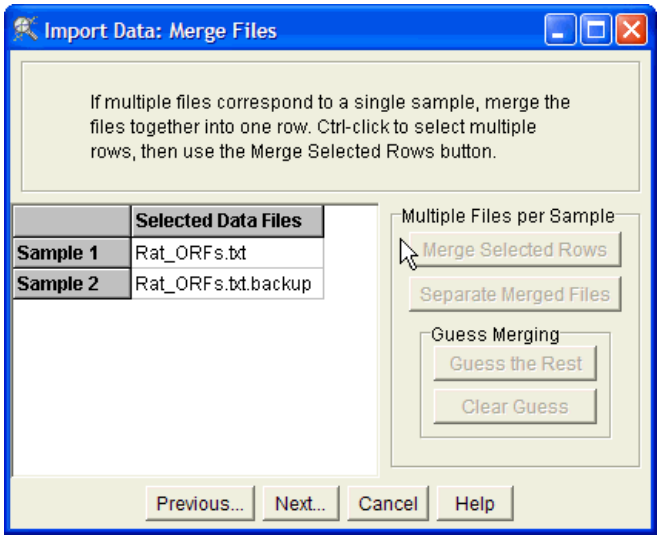


Figure 5-4 Import Data: Merge Files Window

To merge files:

1. Select all the files that represent the same sample.
2. Click the **Merge Selected Rows** button.
   - Use Ctrl+click to select multiple files in non-adjacent rows.
   - You can also drag a file from one row and drop it in another to merge those two rows.
   - To separate the files, select a row, and then click the **Separate Merged Files** button.
3. To enable GeneSpring to try to match the pattern set by the names of the files you have already merged:
   - Click **Guess the Rest** button.

     GeneSpring tries to guess which samples are subchips by examining the file names. For example, if you selected the names *129Amygdala(1a)subA.txt* and *129Amygdala(1a)subB.txt*, it is clear that these two files are subchips. The first part of the file names are similar. Further, the use of “subA” and “subB” in the file names designate them as subchips.

     Choosing appropriate filenames like this makes it easy to import these kinds of file into GeneSpring. Silicon Genetics recommends that you use a similar file naming scheme if you have subchips.
   - If the guesses are incorrect, click the **Clear Guesses** button.

4. Click the **Next** button.
5. Go to "Importing Sample Attributes" on page 5-19.

Managing Data Files with Extra Genes

If your data file contains expression values for genes that are not part of the selected genome, the Extend Genome window opens (Figure 5-5). From this window, you can specify whether or not to add those genes to the genome.

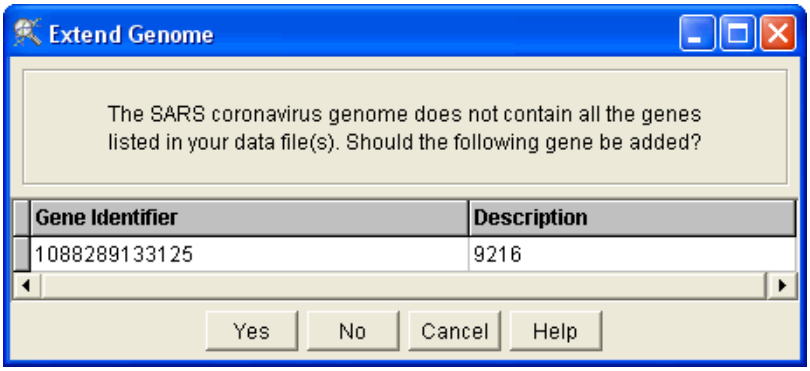


Figure 5-5 Extend Genome Window

If you choose not to add the genes to the genome, the expression values for the genes will not be imported. However, the expression values will be used in the calculation for the normalizations and this may affect the expression value outcome. To ensure more reliable normalizations, Silicon Genetics recommends that you do not exclude genes from the genome in this manner.

To add genes from the data file to the genome:

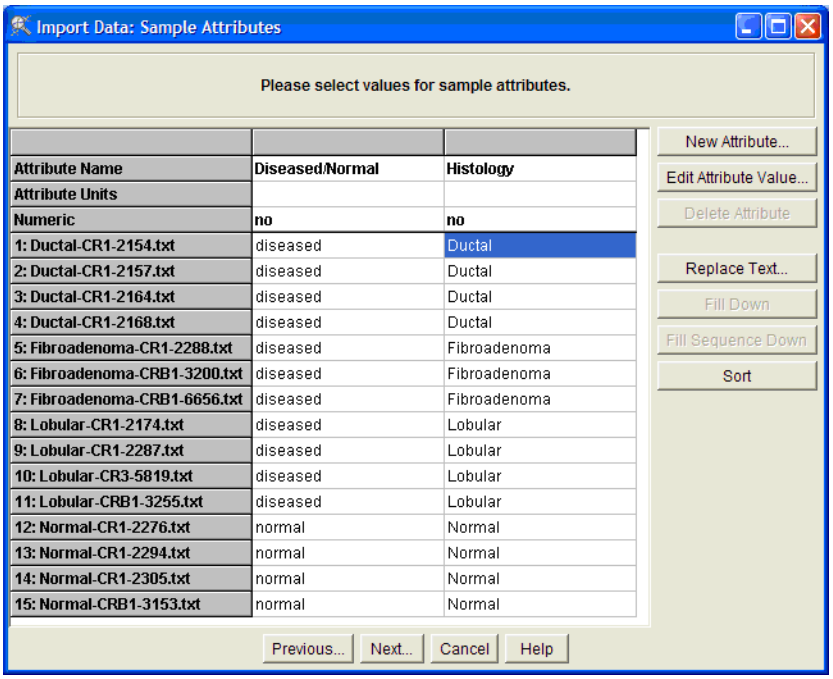
1. Do one of the following:
   - Click the **Yes** button to add the genes.

     Note: If you cancel the data loading process later, the genes are still part of the selected genome.
   - Click the **No** button, to skip the process.
2. Go to "Importing Sample Attributes" on page 5-19.

Importing Sample Attributes

If you defined any required or recommended attributes, the Import Data: Sample Attributes window opens (Figure 5-6). This window lets you enter required attribute information before proceeding. You can also add recommended and optional attribute information.

Note: Some attribute values may have been imported already, depending on the format used.



Attribute Name	Diseased/Normal	Histology
Attribute Units		
Numeric	no	no
1: Ductal-CR1-2154.txt	diseased	Ductal
2: Ductal-CR1-2157.txt	diseased	Ductal
3: Ductal-CR1-2164.txt	diseased	Ductal
4: Ductal-CR1-2168.txt	diseased	Ductal
5: Fibroadenoma-CR1-2288.txt	diseased	Fibroadenoma
6: Fibroadenoma-CRB1-3200.txt	diseased	Fibroadenoma
7: Fibroadenoma-CRB1-6656.txt	diseased	Fibroadenoma
8: Lobular-CR1-2174.txt	diseased	Lobular
9: Lobular-CR1-2287.txt	diseased	Lobular
10: Lobular-CR3-5819.txt	diseased	Lobular
11: Lobular-CRB1-3255.txt	diseased	Lobular
12: Normal-CR1-2276.txt	normal	Normal
13: Normal-CR1-2294.txt	normal	Normal
14: Normal-CR1-2305.txt	normal	Normal
15: Normal-CRB1-3153.txt	normal	Normal

Figure 5-6 Import Data: Sample Attributes Window

The Import Data: Sample Attributes window contains the following elements:

- **Columns—**Displays the annotation columns imported with the data file, or the columns you created.
- **New Attribute—**Displays the New Attribute window and lets you select a standard attribute, or create a custom one.
- **Edit Attribute Value—**Displays the Edit Attribute Value window and lets you edit the selected attribute value.
- **Replace Text**—Displays the standard Replace window which allows you to search and replace text
- **Fill Down**—Fills in all the selected cells with the text from the top selected cell. This button is only enabled when a single column is selected or when two or more cells in the column are selected. GeneSpring will fill all cells below the top cell with the values from the top cell. This tool enables you to quickly copy the contents from one cell to all cells below it and saves you time in entering information.

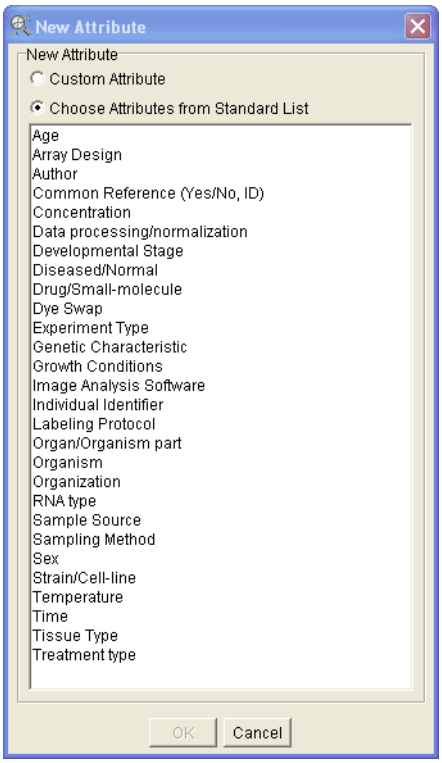
- **Fill Sequence Down—**Fills down a sequence of numbers based on cells that contain the values you want to expand. To fill down a sequence of numbers, select three or more cells, where the first two contain the sequence you want to expand. GeneSpring will try to determine what the spacing between the numbers should be from the first two sequences. If the spacing is 1 (1,2) the sequence will continue with spacing of 1 (1,2,3,4,5, etc.), if the spacing is 10 (10,20) the spacing will be 10 (10,20,30,40,50, etc.).
- **Sort**—Sorts the genes in ascending or descending order based on the entries in the selected column.

The Import Data: Sample Attributes window is similar to the Sample Manager window, For more information on the functions that appear in this window, go to "Working with Sample Attributes" on page 5-58.

To import sample attributes:

1. Click the **New Attribute** button.
2. To select an attribute with a standard value, choose an option from the menu.
3. To turn a pull-down menu into a text field, select **Other** and enter the attribute value.

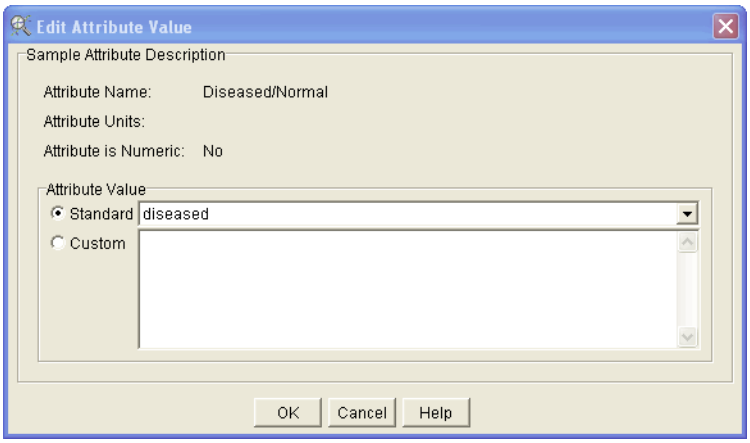
   The New Attribute window opens. The required attributes are colored in yellow and must be filled in.



4. Do one of the following:
   - To add a standard attribute, select it from the menu and click the **OK** button.
   - To add a custom attribute, select the **Custom** option, create the new attribute, and then click the **OK** button.

   A new column appears in the Sample Attributes window.
5. To edit the value of an attribute, click the **Edit Attribute Value** button.

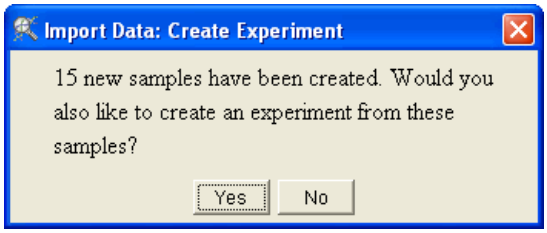
   The Edit Attribute Value window opens.



6. Do one of the following:
   - To add a standard attribute value, select it from the menu and click the **OK** button.
   - To add a custom attribute value, select the **Custom** option, create the new attribute, and then click the **OK** button.

   A new column appears in the Sample Attributes window.
7. When you are done adding attributes, click the **Next** button.

   The Please Wait While Creating Samples window opens. After the samples have been imported, the Import Data: Create Experiment window opens.


Import Data: Create Experiment
15 new samples have been created. Would you also like to create an experiment from these samples?
Yes
No

At this point, your new samples have been saved. You can either create an experiment using the new samples, or stop here.

8. Do one of the following:

   - To stop here, click the **No** button.

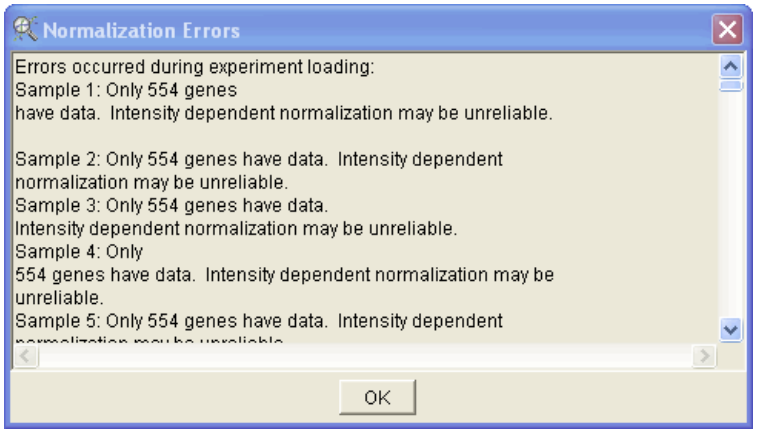
     The imported data are saved, but a new experiment file is not created. The data are saved as samples. The Sample Inspector displays each of the new samples. You can create experiments from these data later by selecting **Experiments > Create New Experiment**.

   - To create a new experiment, click the **Yes** button.

     The Please Wait While Creating Experiment window opens.

9. Do one of the following:

   - If the Normalization Errors window opens, review the information in the window and click the **OK** button to continue.You can renormalize your data at another time. Go to Chapter 10 , “Normalizing Data” for more information.


Normalization Errors
Errors occurred during experiment loading:
Sample 1: Only 554 genes
have data. Intensity dependent normalization may be unreliable.
Sample 2: Only 554 genes have data. Intensity dependent
normalization may be unreliable.
Sample 3: Only 554 genes have data.
Intensity dependent normalization may be unreliable.
Sample 4: Only
554 genes have data. Intensity dependent normalization may be
unreliable.
Sample 5: Only 554 genes have data. Intensity dependent
OK

   - When the Choose Experiment Name window opens, you can save the experiment. Go to “Saving New Experiments” on page 5-23 to continue.

Saving New Experiments

The Save New Experiment window (Figure 5-7) prompts you to enter a name and save the experiment.

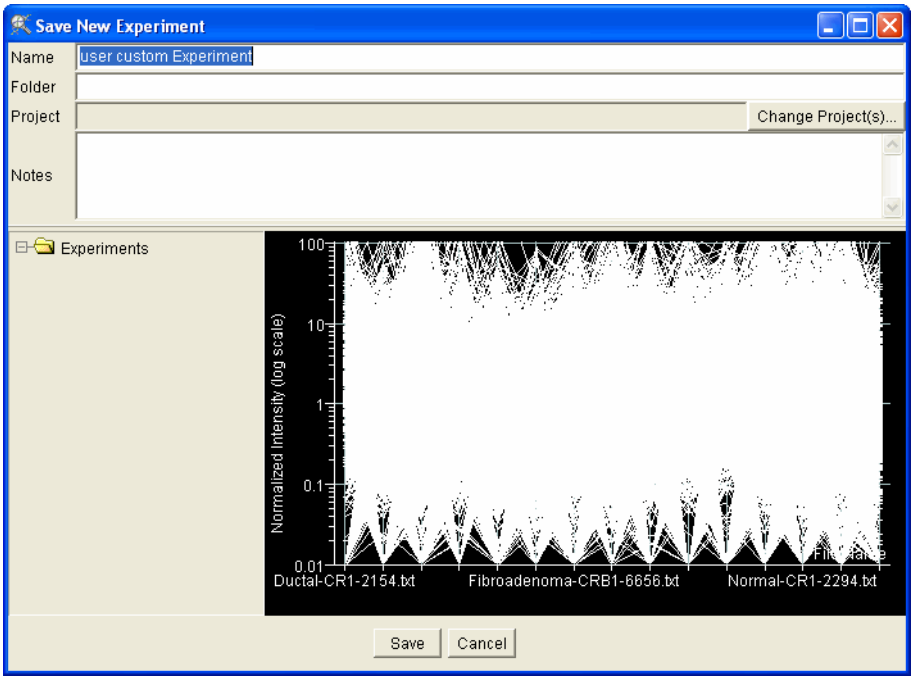
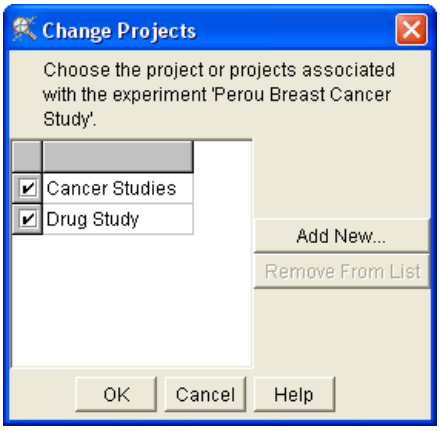


Figure 5-7 Save New Experiment Window

To save a new experiment:

1. In the **Name** field, enter a name for the experiment. Be sure to choose a descriptive name that you will remember later.
2. To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.
3. To save in a new subfolder, navigate to the parent folder and enter a name for the new folder in the **Folder** field.
4. In the **Notes** field, enter any descriptive information about the experiment you want.
5. To assign the experiment to a project, click the **Change Projects** button.

The Change Projects window opens.



6. Do one of the following:
   - To create a new project name, do the following:
     a. Click the **Add New** button.

        The Add New window opens.

     b. In the **Project Name** field, enter a name for the project.

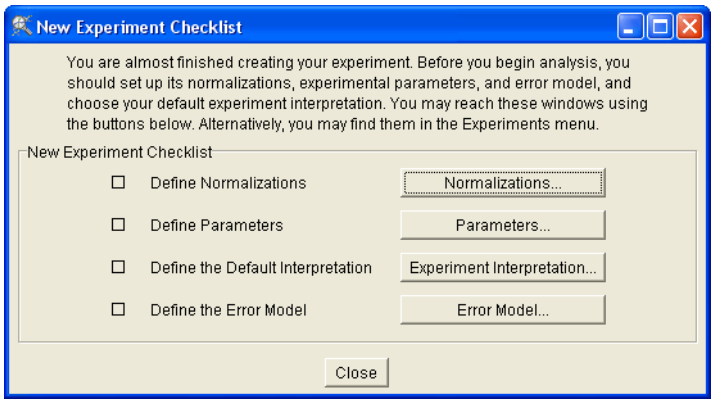
        All data objects in the project will be associated with this project name.

     c. Click the **OK** button.
   - To assign a previously created project name, do the following:
     a. Select the project name(s) you want.

        You can assign data objects to more than one project.

     b. Click the **OK** button.
7. Click the **Save** button.

   The New Experiment Checklist window opens.


New Experiment Checklist
You are almost finished creating your experiment. Before you begin analysis, you should set up its normalizations, experimental parameters, and error model, and choose your default experiment interpretation. You may reach these windows using the buttons below. Alternatively, you may find them in the Experiments menu.
New Experiment Checklist
Define Normalizations
Normalizations...
Define Parameters
Parameters...
Define the Default Interpretation
Experiment Interpretation...
Define the Error Model
Error Model...
Close

At this point, you can examine and change your normalizations, interpretations, and parameters. The New Experiment Checklist will guide you through the steps you need to complete to finalize the creation of the experiment.

Silicon Genetics recommends that you follow all steps in the checklist in order. Using the checklist will ensure that the experiment is set up correctly, which can lead to biological conclusions more easily.

8. Do one or more of the following:

   - To define or edit normalizations, click the **Normalizations** button

     For information on defining normalizations, see "Using the Default Normalizations" on page 5-35.

   - To define or edit parameters, click the **Parameters** button.

     For information on defining parameters, see "Setting Up Experiment Parameters" on page 5-37.

   - To define or edit default interpretations, click the **Experiment Interpretation** button.

     For information on defining default interpretations, see "Setting Up Experiment Interpretations" on page 5-65.

   - To make these changes later, click the **Close** button.

     You can load the experiment another time and select **Experiment Normalizations**, **Change Experiment Parameters**, or **Change Experiment Interpretation** from the Experiments menu in the main GeneSpring window.

Creating New Experiments

In addition to loading expression data from files, as described in "Loading Experiments" on page 5-3, you can create a new experiment using data previously loaded into Genespring. This data can exist on your local system, on a Signet server, or on both. GeneSpring provides a variety of filters to make it easy to select the appropriate samples for your experiment.

Create New Experiment Window

The Create New Experiment window (Figure 5-8) provides tools for filtering sample data, editing sample attributes, and creating an experiment from samples.

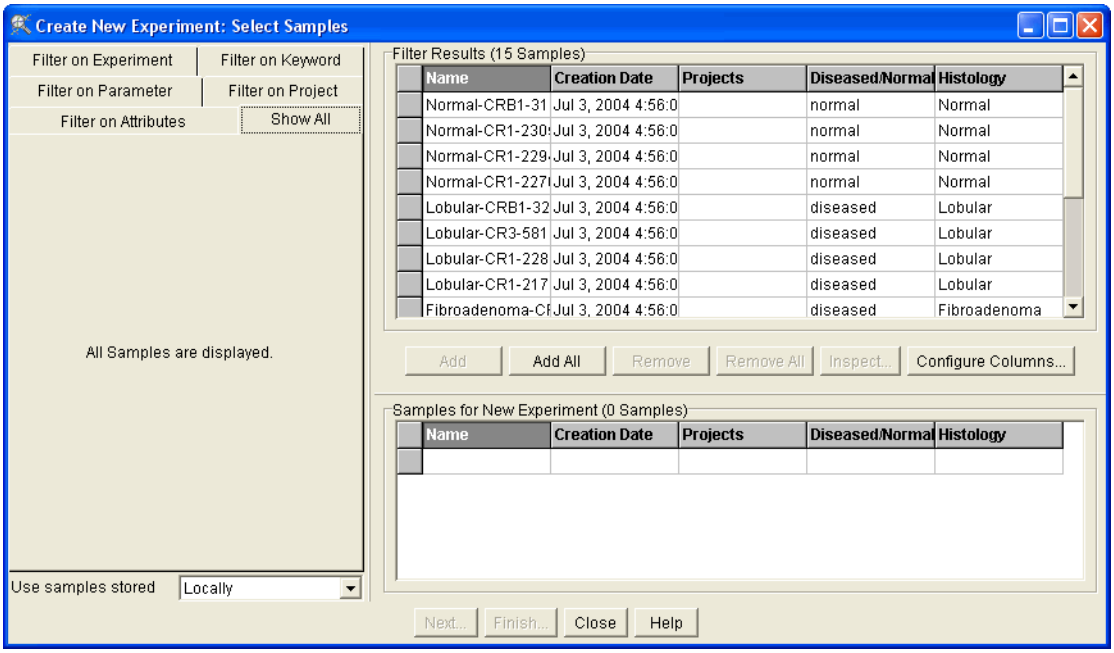


Figure 5-8 Create New Experiment Window

This section describes the elements that appear in the Create New Experiment window.

Note: This window is very similar to the Sample Manager window. For information on related functions not described in this section, go to "Sample Manager Window" on page 5-45.

Filter Methods

The left side of the window contains a tab for each filtering method. The available methods are:

- **Show All**—Displays all available samples without applying a filter.
- **Filter on Experiment**—Displays samples associated with a particular experiment.
- **Filter on Attributes**—Displays samples based on selected attributes.
- **Filter on Keyword**—Displays samples containing a keyword.
- **Filter on Parameter**—Displays samples based on the parameter values of experiments containing them.

Click on the appropriate tab to view the options for that filtering method.

Sample Lists

The right portion of the window contains two sample lists (Figure 5-9). The upper list contains all of the samples resulting from the current filtering method. The lower list contains the samples you have selected.

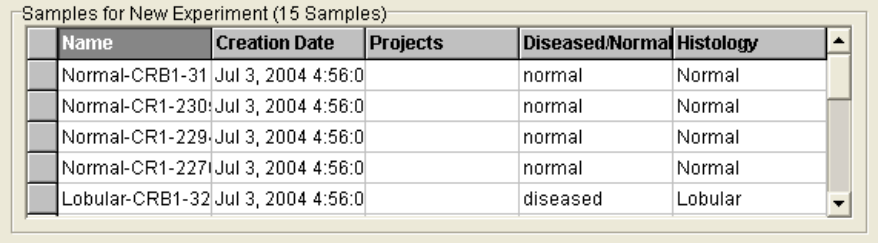


Figure 5-9 Create New Experiment: Selected Samples

Filter Results Buttons

- **Add**—Adds a selected sample in the Filter Results list to the Selected Samples list.
- **Add All**—Adds all samples in the Filter Results list to the Selected Samples list.
- **Remove**—Removes a selected sample from the Selected Samples list.
- **Remove All**—Removes all samples from the Selected Samples list.
- **Inspect—**Shows the selected sample in the Sample Inspector. For more information on the Sample Inspector, see "Using the Sample Inspector" on page 7-19.
- **Configure Columns**—Selects which columns to display in the sample lists. The available choices are:
  - Sample Name
  - Identifier
  - Authors
  - Creation Date
  - Upload Date
  - Research Group
  - Owner
  - Organization
  - Application
  - Projects
  - Sample Attributes
  - Experiment Parameters

Selected Samples Buttons

- **Create Experiment**—Creates a new experiment from the samples in the Selected Samples list.
- **Edit Attributes**—Edits the attributes of the highlighted samples in the Selected Samples list.
- **Assign Projects**—Assign elected data objects to projects.
- **Publish to Signet**—Publishes all samples in the Selected Samples list to Signet.
- **Copy from Signet**—Makes a local copy of all Signet samples in the Selected Samples list.

- **Change Owner**—Changes the owner of the samples.
- **Delete**—Deletes the samples in the Selected Samples list.

Creating New Experiments

To create a new experiment:

1. Select **Experiments > Create New Experiment**.

   The Create New Experiment window opens.

2. Select the samples to include in your experiment.

3. To add a sample, select it in the **Filter Results List**, and then click the **Add** button.

   The sample appears in the **Samples for New Experiment** list.

   - To select multiple samples, hold down the Ctrl key while clicking the samples you want. To add all the samples in the list to your experiment, click the **Add All** button.
   - To view detailed information on a sample, click the **Inspect** button to invoke the Sample Inspector.

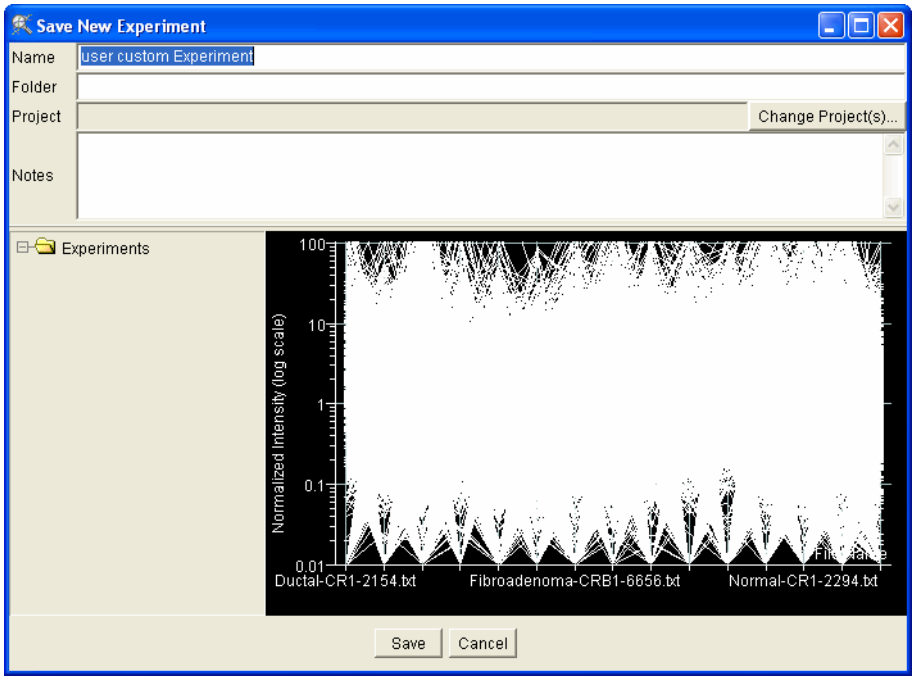
     For more information on the Sample Inspector, see “Using the Sample Inspector” on page 7-19.

4. Do one of the following:

   - If you need to edit parameters or normalizations for this experiment, click the **Next** button.

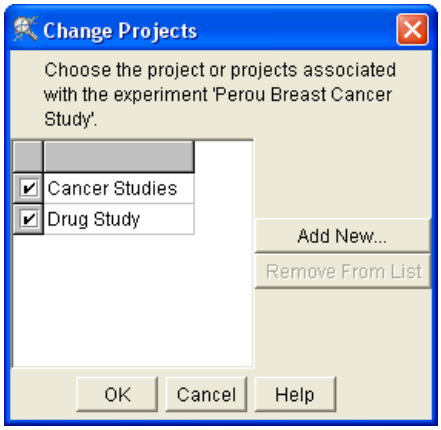
     For detailed information on the Edit Parameters window, see “Setting Up Experiment Parameters” on page 5-37. For detailed information on normalizations, see Chapter 10 , “Normalizing Data”.
   - If you do not need to edit parameters and normalizations, click the **Finish** button.

     The Save New Experiment window opens.



5. In the **Name** field, enter a name for the experiment. Be sure to choose a descriptive name that you will remember later.
6. To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.
7. To save in a new subfolder, navigate to the parent folder and enter a name for the new folder in the **Folder** field.
8. In the **Notes** field, enter any descriptive information about the experiment you want.
9. To assign the experiment to a project, click the **Change Projects** button.

   The Change Projects window opens.



Go to "Working with Projects" on page 7-40 for more information about projects.

10. To create a new project name, do the following:
    a. Click the **Add New** button.

       The Add New window opens.
    b. In the **Project Name** field, enter a name for the project.

       All data objects in the project will be associated with this project name.
    c. Click the **OK** button.
11. To assign a previously created project name, do the following:
    a. Select the project name(s) you want.

       You can assign data objects to more than one project.
    b. Click the **OK** button.
12. When you are done, click the **Save** button.

Copying and Pasting Experiments

You can also use the copy (Ctrl+C) and paste (Ctrl+V) functions to insert a new experiment from the clipboard into GeneSpring.

Preparing to Paste

To prepare to paste, you need data from a spreadsheet or other type of tabular data format. Your data must be organized into three parts, and it *must* be in the following format to correctly paste into GeneSpring:

- Name
- Parameters
- Data

Figure 5-10 shows an example of parameter and values in a spreadsheet.

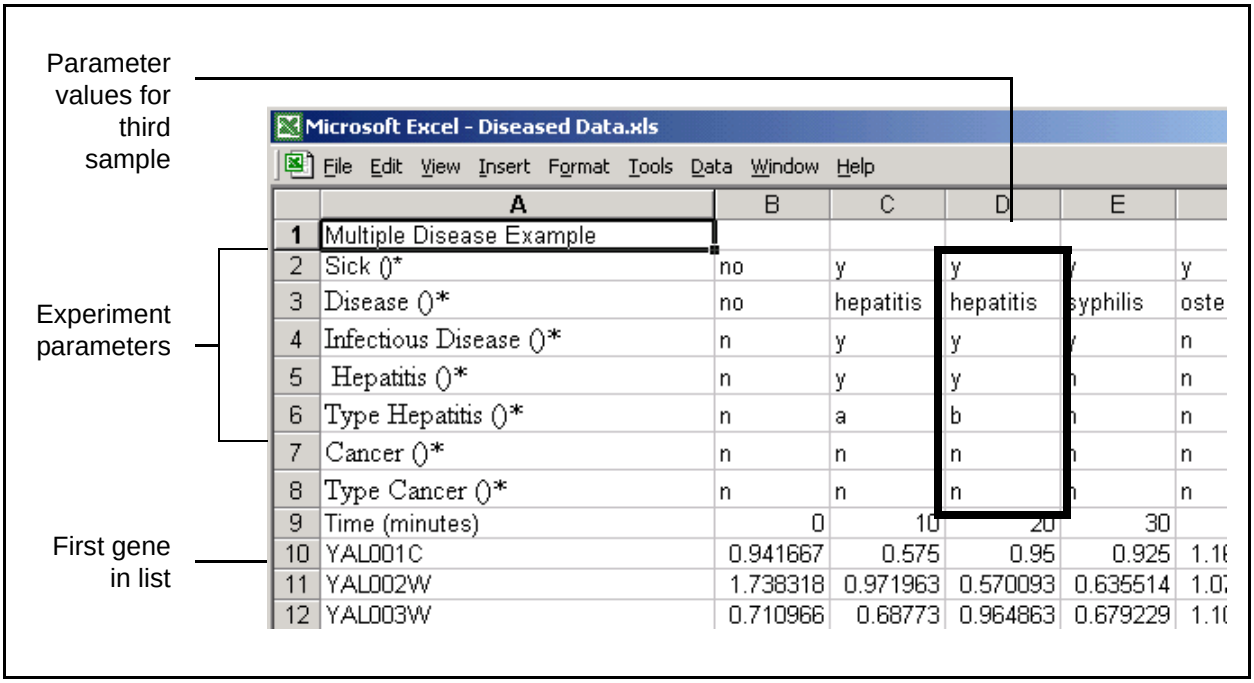

Parameter values for third sample
Experiment parameters
First gene in list
Microsoft Excel - Diseased Data.xls
File Edit View Insert Format Tools Data Window Help
A B C D E
1 Multiple Disease Example
2 Sick ()* no y y y
3 Disease ()* no hepatitis hepatitis syphilis oste
4 Infectious Disease ()* n y y y n
5 Hepatitis ()* n y y n n
6 Type Hepatitis ()* n a b n n
7 Cancer ()* n n n n n
8 Type Cancer ()* n n n n n
9 Time (minutes) 0 10 20 30
10 YAL001C 0.941667 0.575 0.95 0.925
11 YAL002W 1.738318 0.971963 0.570093 0.635514
12 YAL003W 0.710966 0.68773 0.964863 0.679229

Figure 5-10 Example of Parameters and Values in Spreadsheet

The following section describes the format of the spreadsheet.

Name

The first line must be the unique name of the experiment.

Parameters

The second line must be the first parameter. You can have an unlimited number of parameters.

- The first column must contain the parameter name.
- Subsequent columns contain values for the parameter in that sample.

  Each parameter must have units in parentheses in the same column as the name. For example, the parameter “time” should be immediately followed by (minutes). If your parameters have no units you must follow the name with an empty set of parentheses, or GeneSpring does not recognize it as a parameter.

  By default, GeneSpring assumes that the parametric values to follow are numeric and to be displayed in numeric order. If the parametric values for a parameter are non-numeric, enter an asterisk immediately after the unit-indicating parentheses (empty if no units). There must be a space between the right parenthesis and the asterisk. This tells GeneSpring to expect non-numeric parametric values and treat the data appropriately.

The default setting for interpretation of parameters is as a continuous element. See "Continuous Elements" on page 5-70 for details. To have the parameters treated differently, enter the following codes just after the parentheses:

- **S** —Data is interpreted as a non-continuous element, also known as a discrete element. See "Non-Continuous Elements" on page 5-70 for details.
- **C** —Data is colored by the different parametric values assigned automatically by GeneSpring. In Figure 5-11 each column would get a different color as time values 0-160. See "Color Code" on page 5-71 for details.
- **R** —Data is interpreted as a replicate (not shown). See "Hidden Elements" on page 5-71 for details.

You can enter all parameters with the default (with no code after the parentheses) and change the interpretation later from within GeneSpring. See "Setting Up Experiment Interpretations" on page 5-65.

For example, for the parameter `tissue type`, a non-continuous non-numeric parameter, the first column might look like this:

`tissue type() *S`.

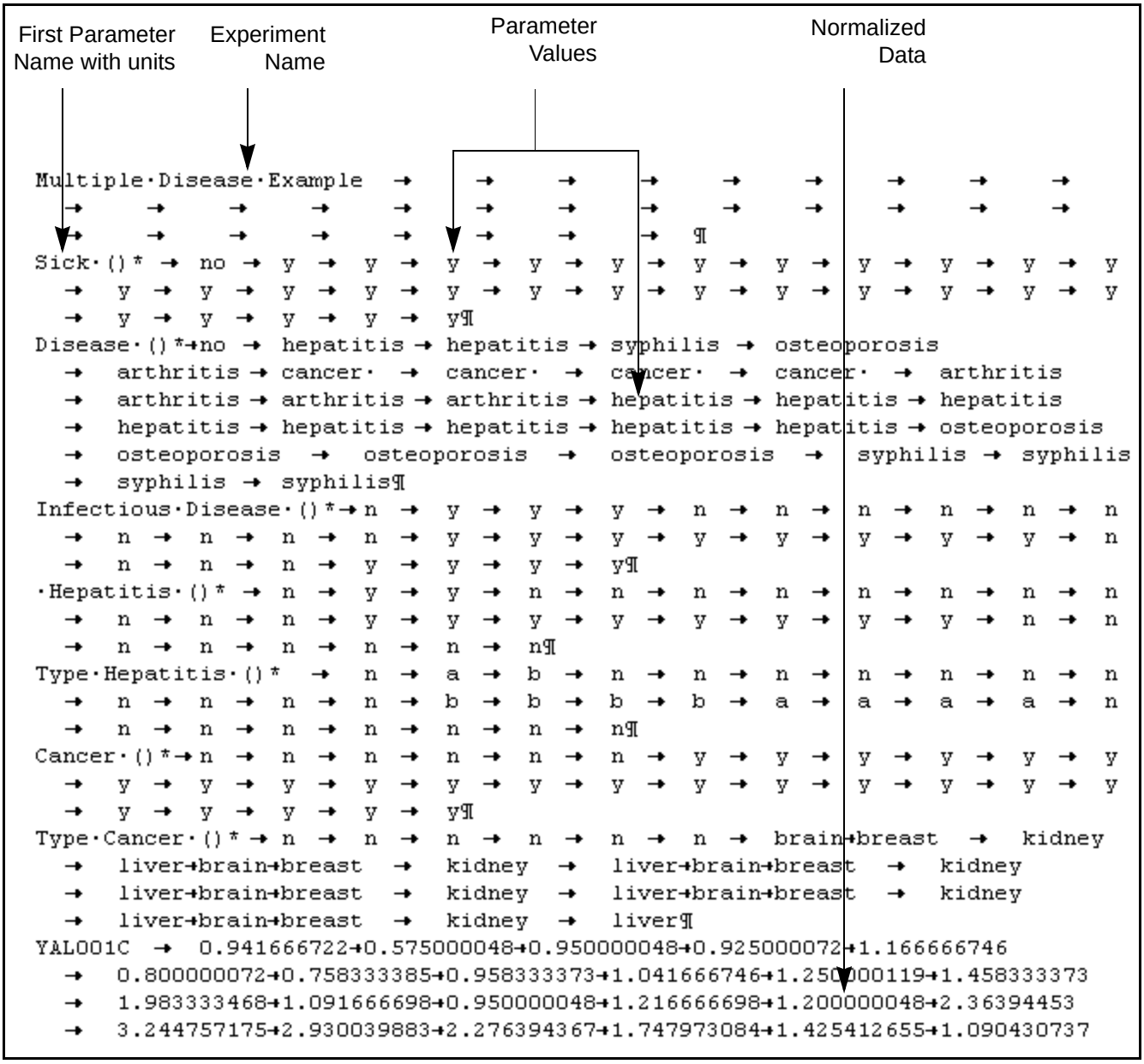
If you have no parameters, enter arbitrary (but meaningful) names so that you can distinguish each sample from those in other columns. In most cases, you would want to import only the normalized data, but if you also want to import the control channel data, ensure that the value for the control channels is "control."

Data

The data should adhere to the following guidelines:

- There can be only one gene per line.
- The name of the gene must be in the first column.
- The normalized data needs to be in the subsequent columns
- If you want to use the Control channel, the data from the control channel needs the be located in the second column to the right of the normalized data.
- For the control channel, the value of the parameter in that column needs to be "control"

The following example (Figure 5-11) shows data points for each sample.




```
Multiple·Disease·Example  →    →    →    →    →    →    →    →    →
  →    →    →    →    →    →    →    →    →    →    →    →    →
  →    →    →    →    →    →    →    →  ¶
Sick·()* → no → y → y → y → y → y → y → y → y → y → y → y → y → y
  → y → y → y → y → y → y → y → y → y → y → y → y → y
  → y → y → y → y ¶
Disease·()*→no → hepatitis → hepatitis → syphilis → osteoporosis
  → arthritis → cancer· → cancer· → cancer· → cancer· → arthritis
  → arthritis → arthritis → arthritis → hepatitis → hepatitis → hepatitis
  → hepatitis → hepatitis → hepatitis → hepatitis → hepatitis → osteoporosis
  → osteoporosis → osteoporosis → osteoporosis → syphilis → syphilis
  → syphilis → syphilis¶
Infectious·Disease·()*→n → y → y → y → n → n → n → n → n → n
  → n → n → n → n → y → y → y → y → y → y → y → n
  → n → n → n → y → y → y → y¶
·Hepatitis·()* → n → y → y → n → n → n → n → n → n → n → n
  → n → n → n → y → y → y → y → y → y → y → n → n
  → n → n → n → n → n → n¶
Type·Hepatitis·()* → n → a → b → n → n → n → n → n → n → n
  → n → n → n → n → b → b → b → b → a → a → a → a → n
  → n → n → n → n → n → n → n¶
Cancer·()*→n → n → n → n → n → n → y → y → y → y → y → y → y
  → y → y → y → y → y → y → y → y → y → y → y → y → y
  → y → y → y → y → y¶
Type·Cancer·()* → n → n → n → n → n → n → brain→breast → kidney
  → liver→brain→breast → kidney → liver→brain→breast → kidney
  → liver→brain→breast → kidney → liver→brain→breast → kidney
  → liver→brain→breast → kidney → liver¶
YAL001C → 0.941666722→0.575000048→0.950000048→0.925000072→1.166666746
  → 0.800000072→0.758333385→0.958333373→1.041666746→1.250000119→1.458333373
  → 1.983333468→1.091666698→0.950000048→1.216666698→1.200000048→2.36394453
  → 3.244757175→2.930039883→2.276394367→1.747973084→1.425412655→1.090430737
```

Figure 5-11 Example of a Correctly Formatted Tab-delimited File

Common Mistakes in Pasting

The following list identifies common mistakes in pasting:

- Forgetting the title
- Not using parentheses
- Not having parameters
- Including extraneous columns
- Forgetting to indicate parameters having non-numeric parametric values with an asterisk (*)
- Using more than one type of decimal marker, or the wrong type for your computer's settings.

Pasting an Experiment into GeneSpring

To paste an experiment into GeneSpring:

1. If you have not done so already, give your experiment a unique name.

   If the name is already in use, GeneSpring appends a number to distinguish it from other experiments of the same name.

2. Select all or part of a properly formatted Excel or tab-delimited file.

3. Click the **Copy** button or press Ctrl-C.

   Note: Some computers have a limit on the amount of data you can place on the clipboard. If you are consistently crashing at the point, you may need a JVM update.

4. Select **Edit > Paste > Paste Experiment**.

   GeneSpring automatically updates the window, regardless of the current display settings. Larger files may take longer to paste, depending on your system.

   When the paste is complete, a new Choose Experiment Name box appears with the current name of the experiment already in the Name text box.

   When you return to the main window, your new experiment is displayed automatically. From here, you can alter the normalizations with **Experiment > Experiment Normalizations** command or the interpretation with the **Experiment > Experiment Interpretation** command.

Copying an Experiment Out of GeneSpring

When you copy an experiment, only data for the currently selected gene list is copied. To copy data for all the genes in the current experiment, click the “All genes” list before you begin to copy. When you paste, the gene list is sorted into the order presented in the Ordered List view.

To copy an experiment out of GeneSpring:

1. In the **Navigator**, select an experiment or a gene list.
2. Select **Edit > Copy > Copy Experiment**.

   Your data is saved to the clipboard.
3. Paste your experiment or gene list into Microsoft Excel or a text editor such as Microsoft Notepad or Microsoft Word.

   Note: The format of the data is identical to the import format when the control channel data is used.

Using the Default Normalizations

Genespring normalizes your new files based on the technology used to create the original data files. If you need to renormalize your files, you can do so using the Normalization window (Figure 5-12).

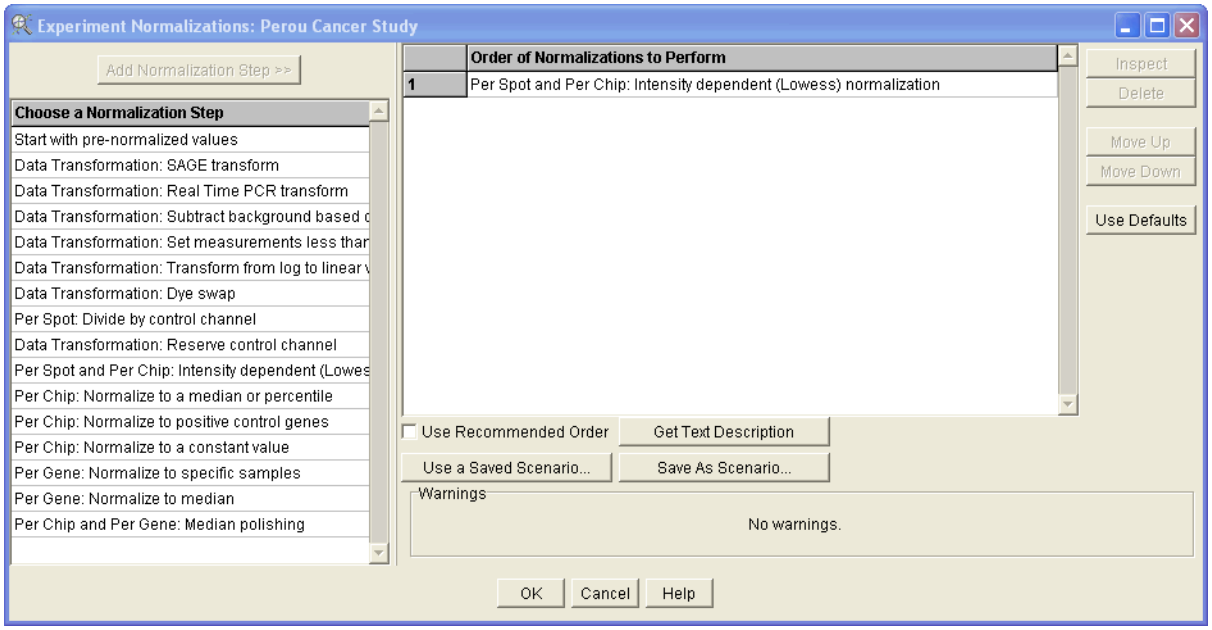

Experiment Normalizations: Perou Cancer Study
Add Normalization Step >>
Choose a Normalization Step
Start with pre-normalized values
Data Transformation: SAGE transform
Data Transformation: Real Time PCR transform
Data Transformation: Subtract background based
Data Transformation: Set measurements less thar
Data Transformation: Transform from log to linear
Data Transformation: Dye swap
Per Spot: Divide by control channel
Data Transformation: Reserve control channel
Per Spot and Per Chip: Intensity dependent (Lowes
Per Chip: Normalize to a median or percentile
Per Chip: Normalize to positive control genes
Per Chip: Normalize to a constant value
Per Gene: Normalize to specific samples
Per Gene: Normalize to median
Per Chip and Per Gene: Median polishing
Order of Normalizations to Perform
1
Per Spot and Per Chip: Intensity dependent (Lowess) normalization
Inspect
Delete
Move Up
Move Down
Use Defaults
Use Recommended Order
Get Text Description
Use a Saved Scenario...
Save As Scenario...
Warnings
No warnings.
OK
Cancel
Help

Figure 5-12 Experiment Normalizations Window

For complete information on normalizations, go to Chapter 10 , “Normalizing Data”.

One-Color Experiments

One-color experiments are experiments that contain expression values determined with the one-color technologies, such as those provided by Affymetrix or Agilent. The standard normalization for one-color data is as follows:

- **Data transformation**—Set measurements less than 0.01 to 0.01.
- **Per Chip**—Normalize to 50th percentile.
- **Options**—Use all flags, never apply background correction.
- **Per Gene—**Normalize to median, cutoff=10 in raw data (if 3+ samples).

Two-Color Experiments

Two-color experiments are experiments that contain expression values that are determined with the two-color technologies, mostly using the Cy3/Cy5 dyes. The normalized value for two-color data are determined by the ratio between the signal and the control channel. The default normalizations are as follows:

- **Per Spot**—Intensity dependent (Lowess) if more than 100 genes per region, divide by control channel if fewer than 100 genes per region. Cutoff = 10 in raw data, 20% of data used for smoothing.
- **Per Chip**—Intensity dependent (Lowess) if more than 1000 genes per chip, divide by control channel if fewer than 1000 genes per chip.
- **Options—**Use background correction if necessary, anything but absent. Cutoff = 10 in raw data.

Pre-Normalized Data

If the data you are importing want to import has been normalized outside of GeneSpring, you can instruct GeneSpring to not perform any normalizations. You can either remove all normalizations from the experiment, or you can transform the data. Data transformation is recommended if the signal and control channel were imported as the signal to control ratio.

Replicates

If you have multiple measurements for the same gene in the same sample, GeneSpring takes the average of the measurements. In addition, GeneSpring saves the minimum and maximum values and display them in the Gene Inspector. See “Handling Repeated Measurements” on page 10-26 for a mathematical explanation of this process.

Remembered Formats

While you cannot edit remembered formats, you can share them. (If you must change a remembered format, you must build a new one.) To share remembered format files, use your favorite browser or file management program to copy the file from Windows:

```
yourlocaldrive:\program files\silicongenetics\
genespring\data\experiment formats\name.expformat
```

Setting Up Experiment Parameters

Experiment parameters are variables that describe the individual samples in a manner that is important for the analysis of the expression data. Experiment parameters are defined for each of the samples of an experiment. This section explains how to set up experiment parameters and includes the following topics:

- Experiment Parameters
- Parameter Values
- Using Parameters
- Parameters and Experiment Interpretations
- Using the Experiment Parameters Window

Experiment Parameters

Experiment parameters are variables that can incorporate many sample values. Generally speaking, when the term parameter is used, it means an experimental parameter. As an example, experiment parameters could be:

- Drug Concentration
- Yeast
- Cancer Outcome
- Replicate

Parameter Values

Parameter values are attributes associated with experiment parameters. For example, the values associated with the list of experiment parameters could be:

- Drug Concentration in ppm
- Strain of Yeast
- Malignant, Benign, or Normal
- Replicate Number

Using Parameters

Experiment parameters are used to group samples with the same parameter value into one group. These groups are called “Conditions” in GeneSpring. Groups can be used to display data in graphs, and they can be used to analyze the variability of genes.

When samples have the same parameter value, they can be grouped into one condition and an average value for that condition can be calculated. A standard deviation can also be calculated for such samples. The standard deviation is used in many of the statistical analysis tools in GeneSpring. By giving samples the same parameter value, you effectively define those samples to be replicate measurements for that particular condition.

For example, if you have five biological or technical replicates for each condition and assign each of the five replicates the same parameter value, GeneSpring will automatically group these samples together and treat them as one condition. Thus, GeneSpring would use one expression value for the gene in that condition, instead of five separate values.

Parameters and Experiment Interpretations

Gene Expression analysis studies usually try to determine a group of genes that are differentially expressed when the expression values between two or more groups (conditions) are compared. The definition of those groups is done through the definition of experiment parameters. GeneSpring uses the idea of "experiment interpretations" to implement the separation of the samples into different groups.

A GeneSpring experiment interpretation is a way to define an experiment. A GeneSpring experiment is a group of related samples, that are separated into one or more groups. The separation into groups is defined by experiment parameters. One GeneSpring experiment can have one or more experiment interpretations. This will allow you to set up your experiment in more than one way.

For example, if you're researching cancer, and you obtained a number of samples you can divide the biological samples in more than one way:

- Cancer outcome: Malignant, Benign, or Normal
- Disease state: Cancerous or Normal
- Concertize: Tumor, Metastatic, or Normal

One way to analyze this data is to determine if there are any genes that are differentially expressed between cancerous and normal samples. Another interesting research question might be to find if there are any genes that are differentially expressed between malignant or benign cancer samples.

In GeneSpring, you can perform any of these kinds of analyses by creating different interpretations for each of the different types of questions. In the example above, you would create three experiment parameters, with their respective values, and create three different experiment interpretations for each of the research questions.

It is also possible to combine two or more parameters and create unique groups (conditions) from them. For example, if you had samples from cell lines, as well as from the patient samples, you could define another experiment parameter, called "CellLine" that would define the cell line of the sample.

With this parameter defined, you could ask questions like, "Are there any genes differentially expressed between cancer and normal cells from cell lines, and are those genes different from those that are differentially expressed between cancer and normal cells from patient samples?". Using parameters in GeneSpring will make answering these questions very easy.

For more information about setting up experiment interpretations, go to "Setting Up Experiment Interpretations" on page 5-65.

Using the Experiment Parameters Window

You can use the Experiment Parameters window (Figure 5-13) to assign parameter names and units, such as time or weight, to your data. You can also use this window to add and delete parameters and rearrange the order of non-numeric parameter values on the horizontal axis. If you set up your file names as described below, your parameter assigning process is automated.

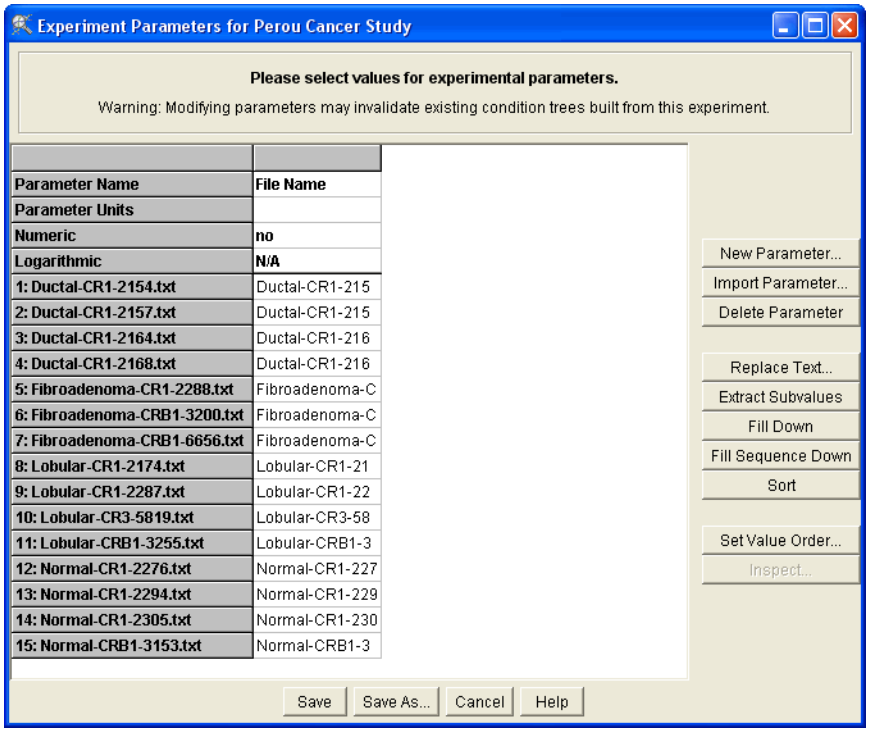


Figure 5-13 Experiment Parameters Window

Importing Parameters

You can import a parameter from another experiment or from a list of sample attributes defined in any of the samples in the current experiment.

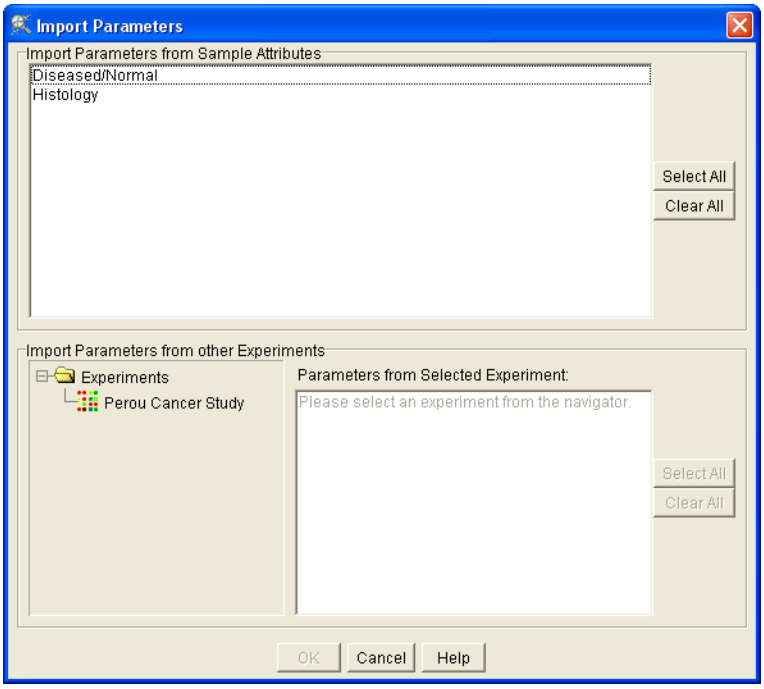
To import a parameter:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.

2. Click the **Import Parameter** button.

The Import Parameters window opens.



3. To select an attribute, do one of the following:
   - Click the attribute you want.

     Only those sample attributes that are associated with the selected samples in the previous window will be available for import.
   - To select all attributes in this list, click the **Select All** button.
   - To clear your selections, click the **Clear All** button.
4. To import parameters from another experiment, select the experiment you want.

   The parameters associated with that experiment appear in the **Parameters from Selected Experiment** list. Only experiments that contain the selected samples will be shown in the experiment selection window.
5. Select the parameters you want.
6. Click the **OK** button.

   The Experiment Parameters window opens. A new column appears for each parameter you imported.

Creating New Parameters

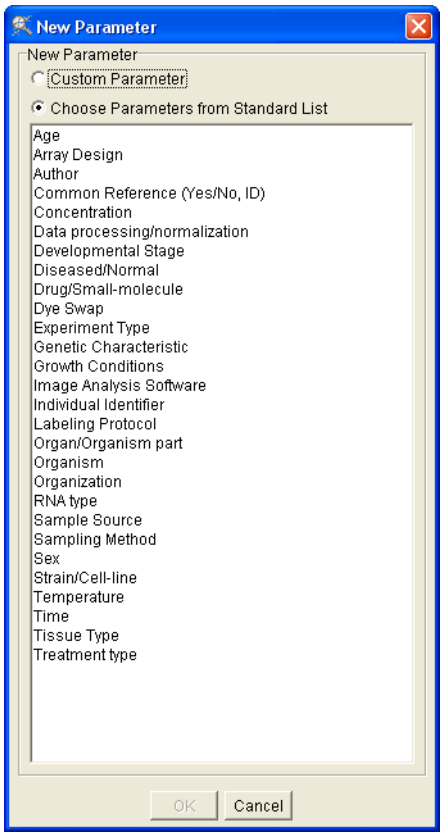
To create a new parameter:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.

2. Click the **New Parameter** button.

   The New Parameter window opens.



3. Do one of the following:

   - To add a standard parameter, select it from the menu, and then click the **OK** button.
   - To add a custom parameter, select the **Custom** option, and then click the **OK** button.

   A new column appears in the Experiment Parameters window.

4. Fill in the **Parameter Name** and **Parameter Units** in the new column (if applicable).

5. In the **Numeric** and **Logarithmic** rows, select **Yes** or **No** from the menus. (Click in a cell in either row to make the menu appear.) You can also paste data in the **Sample** cells.

6. Fill in the parameter values for each sample.

7. Do one of the following:
   - To change the parameters in your current experiment, click the **Save** button.
   - To save this parameter set-up as a new experiment, click the **Save As** button.

Copying and Pasting Data

You can paste in columns of information by clicking the cells of the Sample section. For example, if you have an Excel spreadsheet and want to copy and paste a column from it, you could copy a large section of column and paste it into the new column. You can also copy information out. You can only add columns (parameters and parameter values), you cannot add rows (samples) into this table.

If you have multiple columns you want to copy into the Experiment Parameters window, you should first add as many columns as you want and then copy the contents.

You can also copy the parameter name, unit numeric and Log status. In the spreadsheet, create the columns you want to add and add 4 lines to the top of the columns where you add the Parameter Name, the parameter units, the numeric state (yes/no) the logarithmic state (yes/no) and paste the complete contents by selecting the Parameter name cell in the parameter window and select **Paste** from the short-cut menu.

Deleting Parameters

To delete a parameter:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.
2. Click the gray bar above the column you want to delete.
3. Click the **Delete Parameter** button.

Replacing Text

To replace text:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.
2. Select the entries you want to change
3. Click the **Replace Text** button.

   The Replace window opens.
4. Enter the replacement text.
5. Select any additional options you want. The choices are:
   - Replace in selected cells only.
   - Replace whole values only.
   - Match case.
6. Click the **OK** button.

Extracting Subvalues

The Extract Subvalues function automates parameter assignment. When you use this function, file names are broken down into sub-values. GeneSpring is programmed to first look for alternating constant fields and variable fields and to make parameters out of the variable fields.

Next it divides the variable fields into groups consisting of uninterrupted stretches of either numbers, letters, or non-alpha-numeric characters and makes parameters out of each of these groups.

To extract subvalues:

1. Create file names based on your parameter values.
2. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.
3. Click the **Extract Subvalues** button.

   GeneSpring extracts the subvalues for you and creates as many column as there are experiment parameters and fills the parameter values with the extracted values.

Filling Down Cells

To fill down cells:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.
2. Click on the top-level cell you want to use as the replacement.
3. Holding down the Shift key and click on the cells underneath those values you would like to fill down.
4. Click the **Fill Down** button.

Filling Down Alphabetic or Numeric Sequences

To fill down alphabetic or numeric sequences:

1. Select **Experiments > Experiment Parameters**.

   The Experiment Parameters window opens.
2. Hold down the Shift key and click on the cells underneath those values you would like replaced with the original sequence.

   To fill down a sequence of numbers, select three or more cells, where the first two cells contain the sequence you want to expand. GeneSpring will try to determine what the spacing between the numbers should be from the first two sequences. If the spacing is 1 (1,2) the sequence will continue with spacing of 1 (1,2,3,4,5, etc.), if the spacing is 10 (10,20) the spacing will be 10 (10,20,30,40,50, etc.).
3. Click the **Fill Sequence Down** button.

Changing the Order of Parameters on the Horizontal Axis

When you are using a non-numerical parameter value as the grouping parameter that defines the conditions, and you plot those conditions in the graph view, the default order for the parameters is alphabetical. If the default order is not appropriate for those values, you can manually change the order of the values in the Parameter Value Order window.

To change the order of parameters:

1. Select **Experiments > Experiment Parameters**.

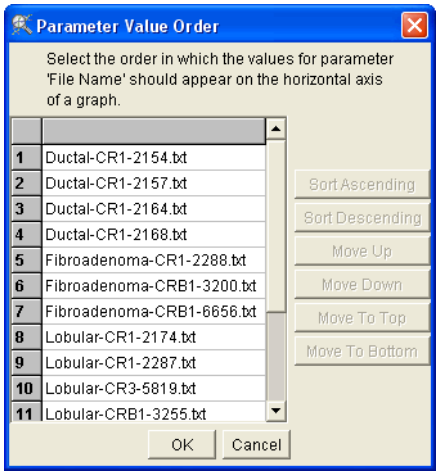
   The Experiment Parameters window opens.

2. Do one of the following:

   - Select an entire column.
   - Select part of a column by clicking in the topmost cell you want to select while holding down the Shift key. GeneSpring selects down the column for you.

3. Click the **Set Value Order** button.

   The Parameter Value Order window opens.



4. To use the **Sort Ascending** or **Sort Descending** buttons, select all the values to be ordered. The main GeneSpring window sorts your parameters according to the new system.

5. To sort manually, select one parameter value and use the Move buttons to arrange the order to your liking.

   Note: You cannot change the order of a parameter defined as numeric.

Working with Samples

In the context of GeneSpring, a sample refers to measurements taken from one or more chips containing a single liquid sample, or the data generated from a biological object placed onto an array or set of arrays.

This section explains how to use the Sample Manager to filter and add samples to your experiments. It covers the following topics:

- Sample Manager Window
- Opening the Sample Manager
- Filtering on Experiments
- Filtering on Parameters
- Filtering on Attributes
- Filtering on Keyword
- Finding Similar Samples
- Viewing Similar Sample Results

Sample Manager Window

The Sample Manager window (Figure 5-14) provides tools for filtering sample data, editing sample attributes, and creating an experiment from samples.

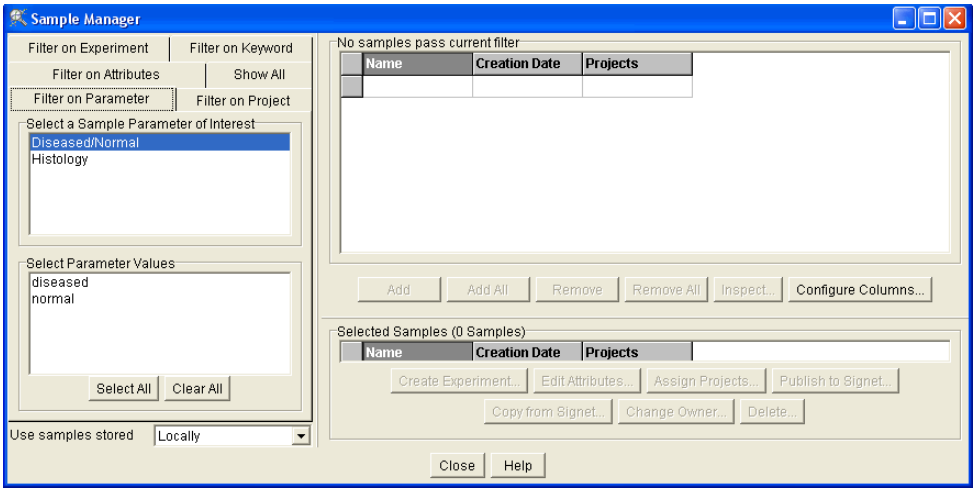


Figure 5-14 Sample Manager Window

The following section describes the elements in this window.

Filtering Methods

The left side of the window contains a tab for each filtering method. The available methods are:

- **Show All**—Displays all available samples without applying a filter
- **Filter on Experiment**—Displays samples associated with a particular experiment
- **Filter on Attributes**—Displays samples based on selected attributes
- **Filter on Keyword**—Displays samples containing a keyword
- **Filter on Parameter**—Displays samples based on the parameter values of experiments containing them

Click on the appropriate tab to view the options for that filtering method.

Sample Lists

The right portion of the window contains two sample lists (Figure 5-15). The upper list contains all of the samples resulting from the current filtering method. The lower list contains the samples you have selected.

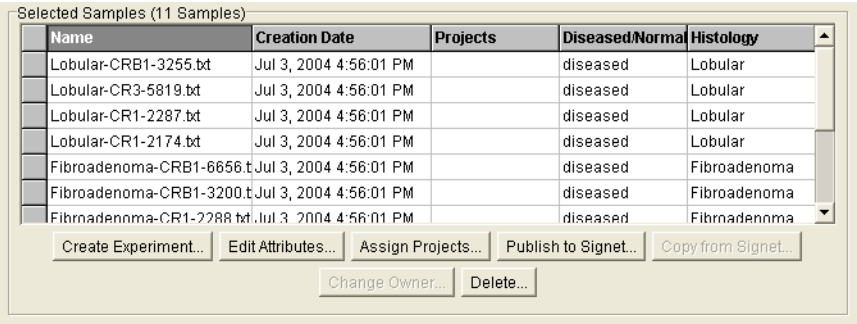


Figure 5-15 Sample Manager Window: Selected Samples

Filter Results Buttons

- **Add**—Adds a selected sample in the Filter Results list to the Selected Samples list.
- **Add All**—Adds all samples in the Filter Results list to the Selected Samples list.
- **Remove**—Removes a selected sample from the Selected Samples list.
- **Remove All**—Removes all samples from the Selected Samples list.
- **Inspect—**Shows the selected sample in the Sample Inspector. For more information on the Sample Inspector, see "Using the Sample Inspector" on page 7-19.
- **Configure Columns**—Selects which columns to display in the sample lists. The available choices are:
  - Sample Name
  - Identifier
  - Authors

- Creation Date
- Upload Date
- Research Group
- Owner
- Organization
- Application
- Projects
- Sample Attributes
- Experiment Parameters

Selected Samples Buttons

- **Create Experiment**—Creates a new experiment from the samples in the Selected Samples list.
- **Edit Attributes**—Edits the attributes of the highlighted samples in the Selected Samples list.
- **Assign Projects**—Assign selected data objects to projects.
- **Publish to Signet**—Publishes all samples in the Selected Samples list to Signet.
- **Copy from Signet**—Makes a local copy of all Signet samples in the Selected Samples list.
- **Change Owner**—Changes the owner of the samples.
- **Delete**—Deletes the samples in the Selected Samples list.

Opening the Sample Manager

To open the Sample Manager:

- Select **Experiments > Sample Manager**.

Filtering Samples

This section explains how to use the Sample Manager to filter samples using different search criteria. These functions include the following:

- Filtering on Experiments
- Filtering on Parameters
- Filtering on Attributes
- Filtering on Keyword
- Filtering on Projects

Filtering on Experiments

The Filtering on Experiment tab (Figure 5-16) lets you find samples associated with a selected experiment.

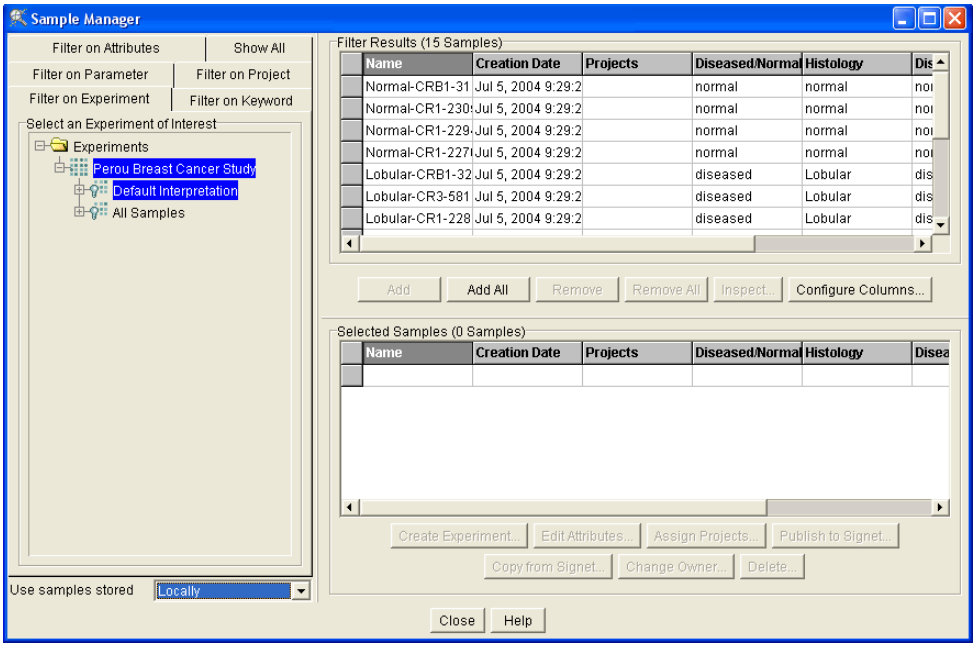


Figure 5-16 Sample Manager Window: Filter on Experiment Tab

To filter on experiments:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

2. Click the **Filter on Experiment** tab.
3. In the **Use samples stored** list, select the location of the samples you want to search.
4. In the **Navigator**, select the **Experiments** folder.
5. Select the experiment you want.

All samples associated with that experiment appear in the Filter Results list.

Filtering on Parameters

The Filtering on Parameter tab (Figure 5-17) filters samples based on a parameter and value range. Parameters are associated with samples that are used in conjunction with an experiment of which they are a member. The samples listed are those contained in any experiment matching the selected parameter and values.

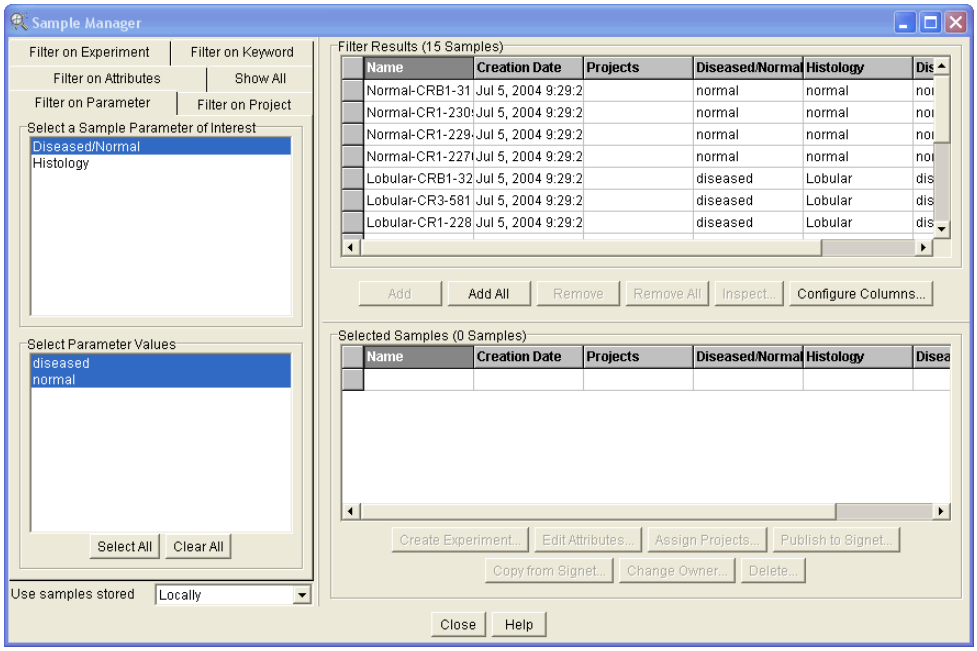


Figure 5-17 Sample Manager Window: Filter on Parameter Tab

To filter on parameters:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

2. Click the **Filter on Parameter** tab.
3. In the **Use samples stored** list, select the location of the samples you want to search.
4. In the **Select a Sample Parameter of Interest** list, select the parameter you want.
5. In the **Select Parameter Values** list, select the parameter value you want:
   - To select all values in the list, click the **Select All** button.
   - To clear your selections click the **Clear All** button.

   The Filter Results list is updated dynamically as you make your selections.
6. To remove a parameter value, click the value again to de-select it.

   This action removes the sample that matched that parameter from the filter results.

Filtering on Attributes

The Filtering on Attributes tab (Figure 5-18) lets you filter samples based on attributes.

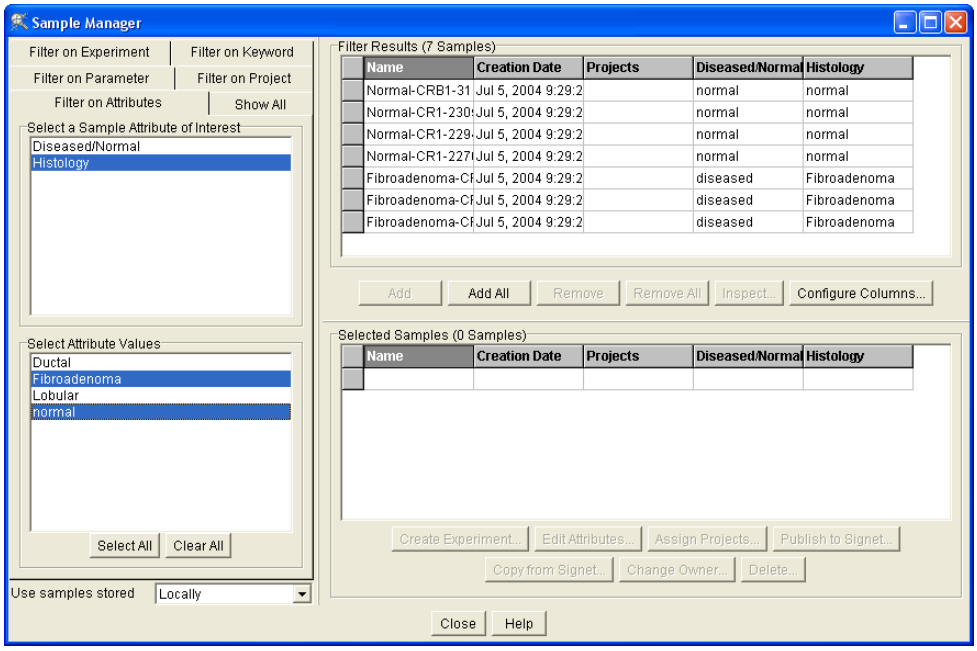


Figure 5-18 Sample Manager Window: Filter on Attributes Tab

Attributes are very similar to parameters, but are associated with individual samples rather than in conjunction with experiments. Attributes can contain sample-specific information that would not be typically used as a basis for analysis. However, by copying the sample attribute as an experiment parameter, the sample attributes can be used as experiment parameters. The following list shows examples of sample attributes:

- Patient biography
- Lab technician
- Date
- Ambient temperature

To filter on attributes:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

2. Click the **Filter on Attributes** tab.
3. In the **Use samples stored** list, select the location of the samples you want to search.
4. In the **Select a Sample Attribute of Interest** list, select the attribute you want.

5. In the **Select Attribute Values** list, select the attribute value you want.
   - To select an attribute value, single-click the value.
   - To deselect an attribute value, single-click the value.
   - To select all values in the list, click the **Select All** button.
   - To clear your selections click the **Clear All** button.

   The Filter Results list is updated dynamically as you make your selections.

Filtering on Keyword

The Filter on Keyword tab (Figure 5-19) lets you filter based on whether a particular keyword appears in any of the specified fields. This method is useful in cases where a given string (such as “cancer”) is sometimes the parameter, sometimes the parameter value, and sometimes is part of the experiment name.

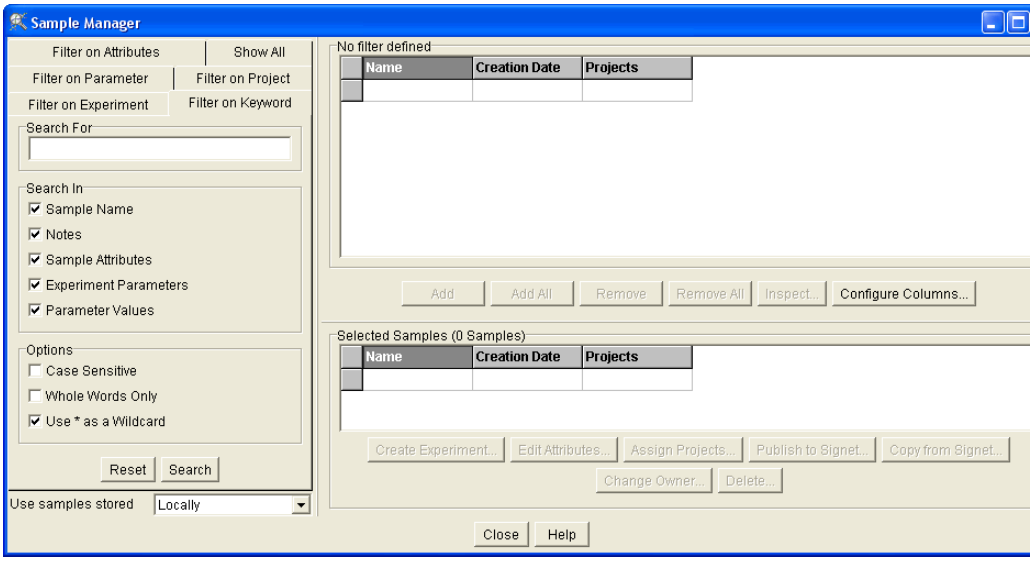


Figure 5-19 Sample Manager Window: Filter on Keyword Tab

To filter on keywords:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

2. Click the **Filter on Keyword** tab.

3. In the **Use samples stored** list, select the location of the samples you want to search.

4. In the **Search For** box, enter the keyword you want.

   This keyword can be a word or a number. It can also contain an asterisk (*) as a wildcard character.

5. In the **Search In** panel, select the fields to search.

6. To exclude a field from the search, click the box to remove the check. You can select as many fields or as few fields as you like. The available choices are:
   - Sample Name
   - Notes
   - Sample Attributes (includes the values of the attributes)
   - Experiment Parameters
   - Parameter Values
7. In the **Options** panel, select any additional search features. You can choose from the following:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
8. Click the **Search** button.

   Any samples matching your search criteria appear in the Filter Results list.

Filtering on Projects

The Filter on Projects tab (Figure 5-20) lets you filter samples based on projects assignments.

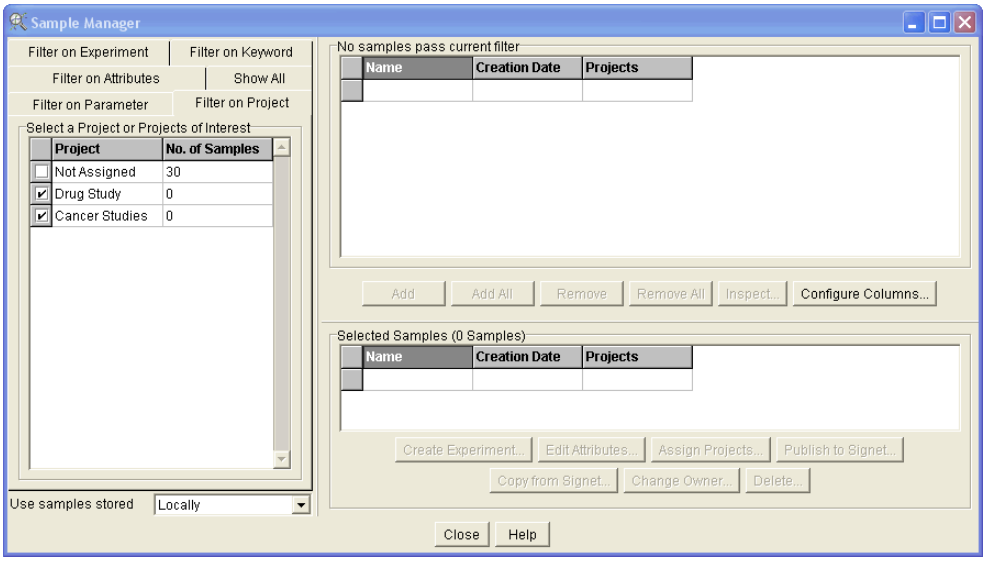


Figure 5-20 Sample Manager Window: Filter on Projects Tab

To filter on projects:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

2. Click the **Filter on Projects** tab.

3. Under the **Select a Project or Projects of Interest** column, select the projects you want.

   Any samples matching your search criteria appear in the Filter Results list.

Finding Similar Samples

The Find Similar Samples tool lets you find samples with genes that show a similar expression profile to a sample you select. This tool will allow you to find other experimental conditions which result in the gene expression profile you see in your sample.

Note that the tool Find Similar Samples does not select the most similar samples. It just grades the pool samples by their similarity to the target sample, and then displays all of them ordered by their grades.

Find Similar Samples and Sample Correlations Tools

The Find Similar Samples tool is similar to the Sample Correlations tool in the Sample Inspector window, but is also different in a number of important aspects. The Sample Correlation window only allows the sample to be compared with other sample in the same experiment, while the Find Similar Samples tool lets you find similar samples in other experiments or samples that are not part of an experiment.

Since samples in other experiments (or outside experiments) are most likely normalized in a different way (or not at all, in the case of individual samples, not part of an experiment), the Find Similar Samples tool will be able to perform some sort of normalization on the fly.

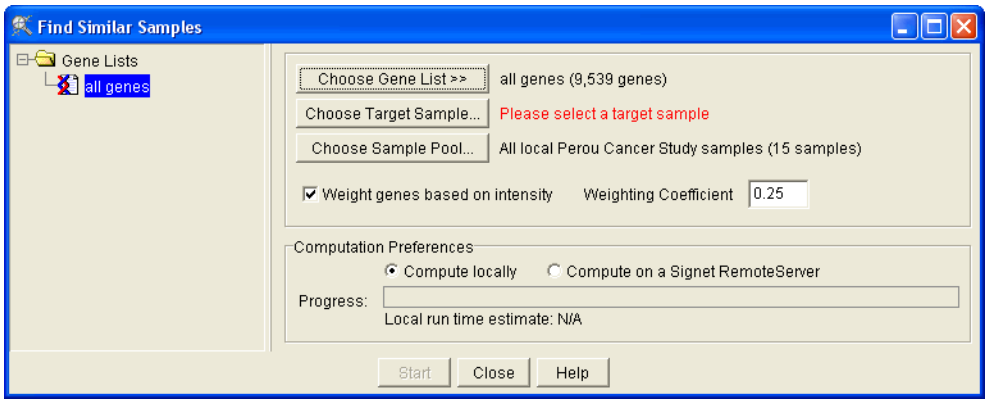
Finding Similar Samples

To find a similar sample:

1. Do one of the following:
   - Open the **Sample Inspector**, select the **Similar Samples** tab, and then click the **Find Similar Samples** button.

     Go to "Using the Sample Inspector" on page 7-19 for more information.
   - Select **Tools > Find Similar Samples**

     The Find Similar Samples window opens.

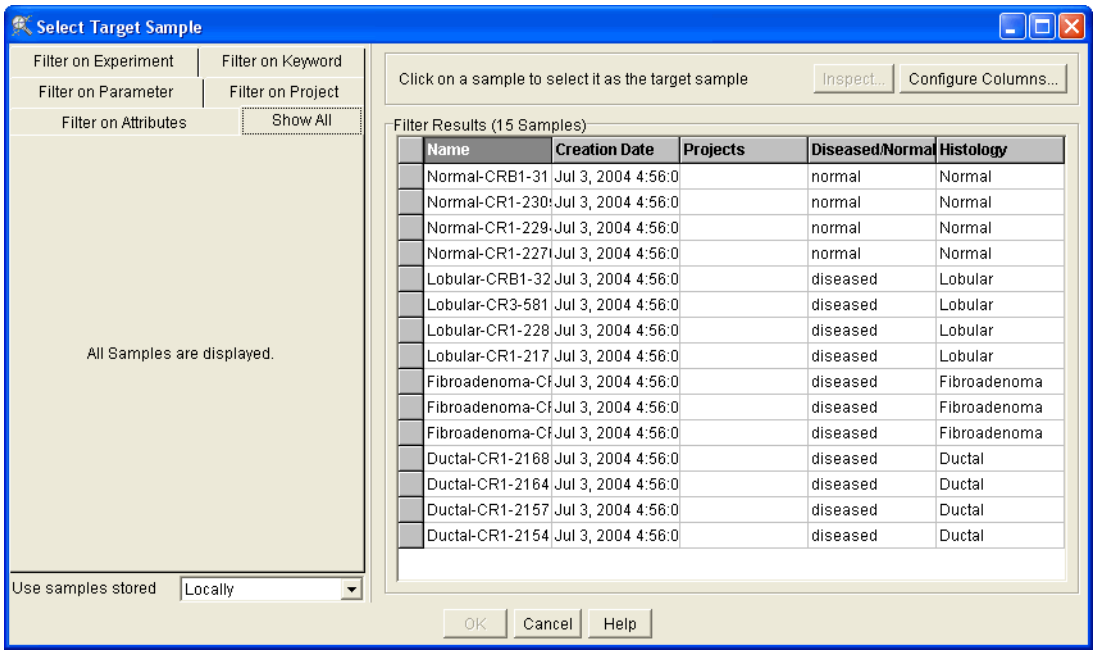


2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List button**.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged "Absent".

3. In the **Navigator**, select the target sample you want to analyze, and then click the **Choose Target Sample** button.

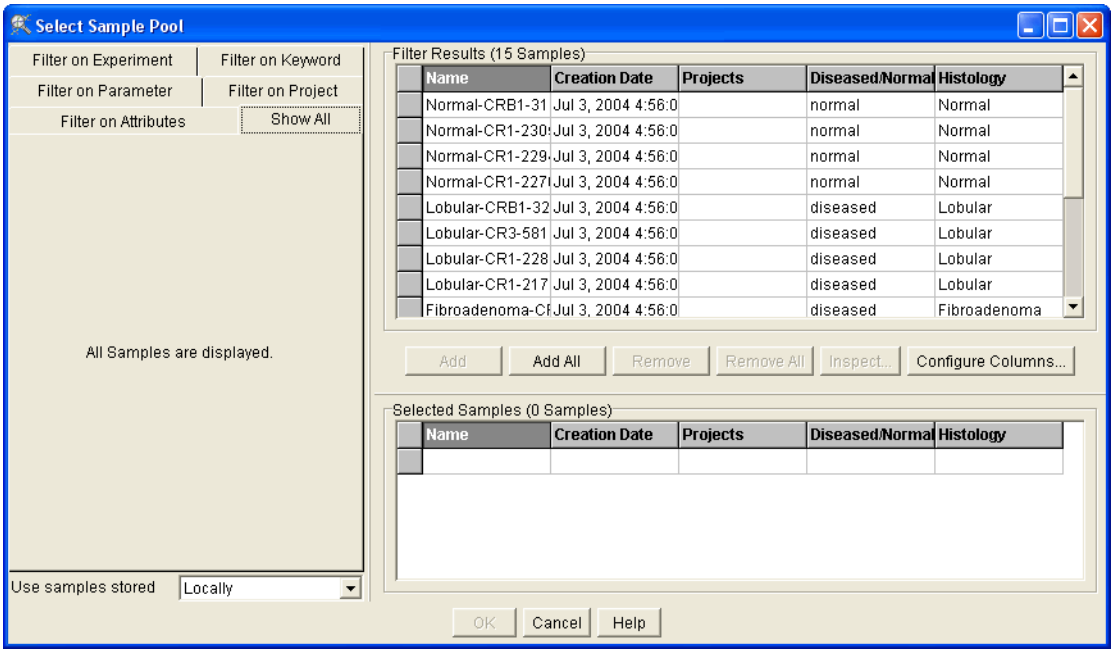
   The Select Target Sample window opens.


Select Target Sample
Filter on Experiment
Filter on Keyword
Filter on Parameter
Filter on Project
Filter on Attributes
Show All
All Samples are displayed.
Use samples stored
Locally
Click on a sample to select it as the target sample
Inspect...
Configure Columns...
Filter Results (15 Samples)
Name
Creation Date
Projects
Diseased/Normal
Histology
Normal-CRB1-31 Jul 3, 2004 4:56:0 normal Normal
Normal-CR1-230 Jul 3, 2004 4:56:0 normal Normal
Normal-CR1-229 Jul 3, 2004 4:56:0 normal Normal
Normal-CR1-227 Jul 3, 2004 4:56:0 normal Normal
Lobular-CRB1-32 Jul 3, 2004 4:56:0 diseased Lobular
Lobular-CR3-581 Jul 3, 2004 4:56:0 diseased Lobular
Lobular-CR1-228 Jul 3, 2004 4:56:0 diseased Lobular
Lobular-CR1-217 Jul 3, 2004 4:56:0 diseased Lobular
Fibroadenoma-C Jul 3, 2004 4:56:0 diseased Fibroadenoma
Fibroadenoma-C Jul 3, 2004 4:56:0 diseased Fibroadenoma
Fibroadenoma-C Jul 3, 2004 4:56:0 diseased Fibroadenoma
Ductal-CR1-2168 Jul 3, 2004 4:56:0 diseased Ductal
Ductal-CR1-2164 Jul 3, 2004 4:56:0 diseased Ductal
Ductal-CR1-2157 Jul 3, 2004 4:56:0 diseased Ductal
Ductal-CR1-2154 Jul 3, 2004 4:56:0 diseased Ductal
OK
Cancel
Help

   The Select Target Sample window provides functions that are similar to the Sample Manager, except that you can select only one sample. For details on using these functions, go to "Sample Manager Window" on page 5-45.

4. To specify the samples to search, click **Choose Sample Pool**.

   The Select Sample Pool window opens.



The Select Sample Pool window provides functions that are similar to the Sample Manager. For details on using these functions, go to "Sample Manager Window" on page 5-45.

5. In the **Find Similar Samples** window, change the value in the **Weighting Coefficient** box., if necessary.

   The weighting coefficient ranges from 0 to 1. The default weighting coefficient is 0.25. If you chooses WC = 0, all the weights will be the same. This is just a regular, non-weighted correlation. If you choose a larger weighting coefficient, highly expressed genes will have more influence on the results.

   The weighting coefficient gives greater emphasis to genes with higher expression value. If the weighting coefficient is positive (and the check box is selected), the genes with the higher raw expression value (or with higher control value, for 2-color data) contribute more. The higher the weighting coefficient, the more significant the contribution of the highly-expressed genes, and the less the contribution of the under-expressed ones.

6. If you do not wish to weight genes based on their control value, clear the **Weight genes based on intensity** check box.

   Selecting this option is equivalent to setting the weighting coefficient to zero. In this case, all valid genes contribute evenly to the correlation calculated.

7. Specify whether to run this process locally, or on a Signet server.

8. Click the **Start** button.

   When the analysis is complete, the Find Similar Samples Results window opens.

9. Go to "Viewing Similar Sample Results" on page 5-57.

Viewing Similar Sample Results

The Find Similar Samples Results window (Figure 5-21) displays the results of the Find Similar Samples operation, both as a bar graph ordered by correlation and a list of samples.

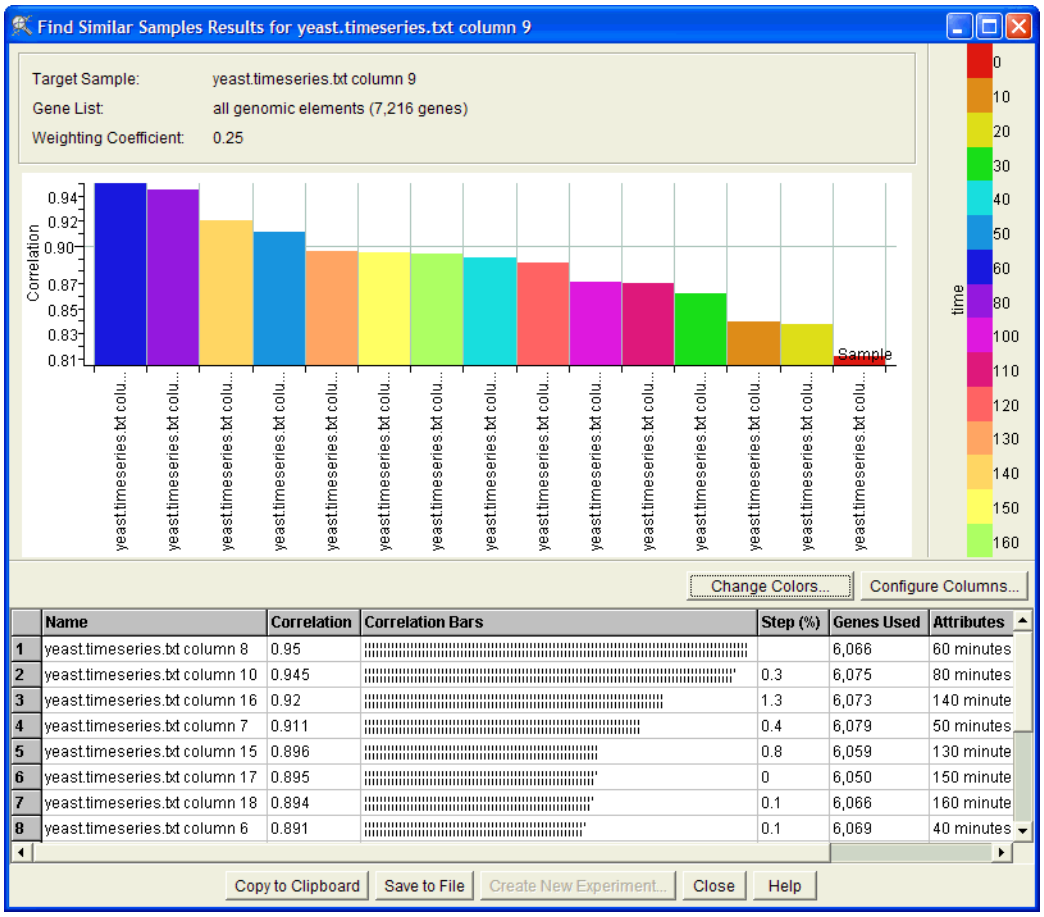


Figure 5-21 Find Similar Samples Results Window

The top portion of this window displays a bar graph of your query results. The lower portion of the window displays a table of the samples returned by the query. By default, the samples are listed in the order of their correlation with the sample to which they were compared.

To view the find similar results:

1. To reorder rows, click a column header.
2. To view that sample in the Sample Inspector, double-click a row.

   The following options are available:

   - **Change Colors**—Specifies the colors used in the bar graph display. You can color the graph using a single solid color (the default), or by a selected sample attribute. If

no sample attributes are defined for the samples, the **Attribute** option does not appear.

- **Configure Columns**—Specifies what information to display in the table of samples.
- **Copy to Clipboard**—Copies the information in the table of samples to the clipboard. This data can be pasted into a text file or a spreadsheet application such as Excel.
- **Save to File**—Saves the information in the table of samples.
- **Create New Experiment**—Creates a new experiment from selected samples. For details, see "Creating New Experiments" on page 5-25.

3. Click the **Close** button when you are done.

Working with Sample Attributes

The Sample Manager is an important part of the experiment creation process, but it can also be used on its own. For example, the Sample Manager lets you filter samples associated with a particular experiment, attribute, or keyword. The Sample Manager also lets you create attributes, edit attribute values, and upload samples to Signet.

This section explains how to use the Sample Manager to manage attributes and values in GeneSpring. It includes the following topics:

- Sample Attributes
- Editing Sample Attributes
- Importing Sample Attributes from Experimental Parameters
- Creating New Sample Attributes
- Deleting Sample Attributes
- Using the Standard Attributes Editor

In addition to managing attributes, the Sample Manager can also be used to filter sample data. Go to "Working with Projects" on page 7-40 for more information.

Sample Attributes

Samples in GeneSpring are equivalent to single hybridization experiments. The sample definitions can contain both information on the biological origin of the sample, as well as information about the hybridization conditions. The information for a sample is stored in sample attributes. These attributes have a name, units, indication if the value is numeric, and the actual value itself.

Attributes can also contain information about a sample, but this information is not necessarily needed to analyze the data. Attributes, like hybridization temperature and operator, usually do not factor in as variables that you want to control for (except for quality assurance purposes) in the expression analysis.

However, samples can have attributes that are required for data analysis, and although this kind of information is usually contained in experiment parameters, sample attributes can be copied into experiment parameters easily.

For more information about selecting and associating attributes with a sample, go to "Attributes and Parameters Tab" on page 7-21.

You can add any number of additional attributes to a sample using the Sample Manager or the Standard Attribute Editor. Additional sets of sample attributes are available for download from the Silicon Genetics web site. The standard set of sample attributes contain attributes that can be used to ensure that your data adheres to the MIAME standard. For more information about the MIAME standard, see http://www.mged.org/Workgroups/MIAME/miame.html.

Editing Sample Attributes

The Edit Attributes window (Figure 5-22) lets you add, edit, or delete sample attributes.

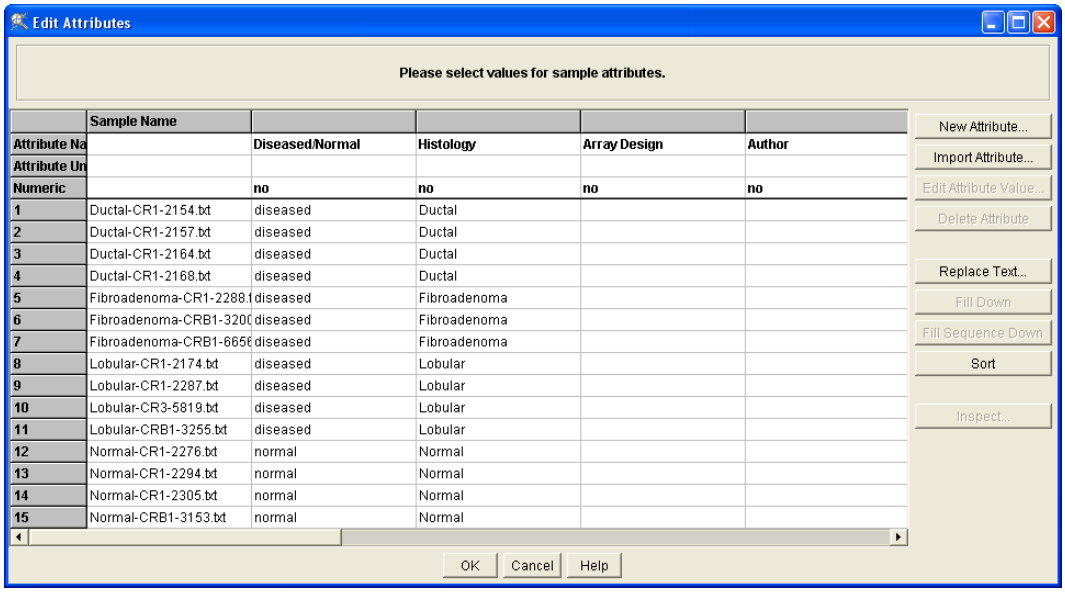


Figure 5-22 Edit Attributes Window

Importing Sample Attributes from Experimental Parameters

You can import an attribute from another experiment or from a list of sample attributes defined in any of the samples in the current experiment. you can also convert a parameter into a sample attribute.

To import a sample attribute from an experiment parameter:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.

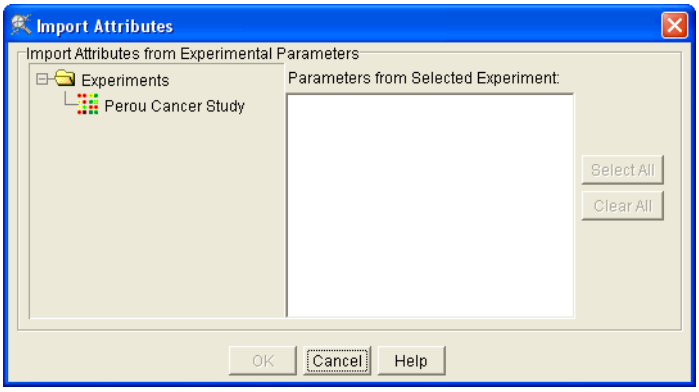
2. Use a filter to select the samples you want.

Go to "Filtering Samples" on page 5-47 for instructions.

3. Click the **Add** button to add the samples to the Selected Samples table.
4. Click the **Edit Attributes** button.

   The Edit Attributes window opens.
5. Click the **Import Attributes** button.

   The Import Attributes window opens.



6. From the **Navigator**, select the attribute you want to import.
7. Do one of the following:
   - To select individual attributes, select the ones you want.
   - To select all the attributes, click **Select All**.
8. Click the **OK** button.

   The Edit Attributes window opens. A new column appears for each attribute you imported from experiment parameters.

Creating New Sample Attributes

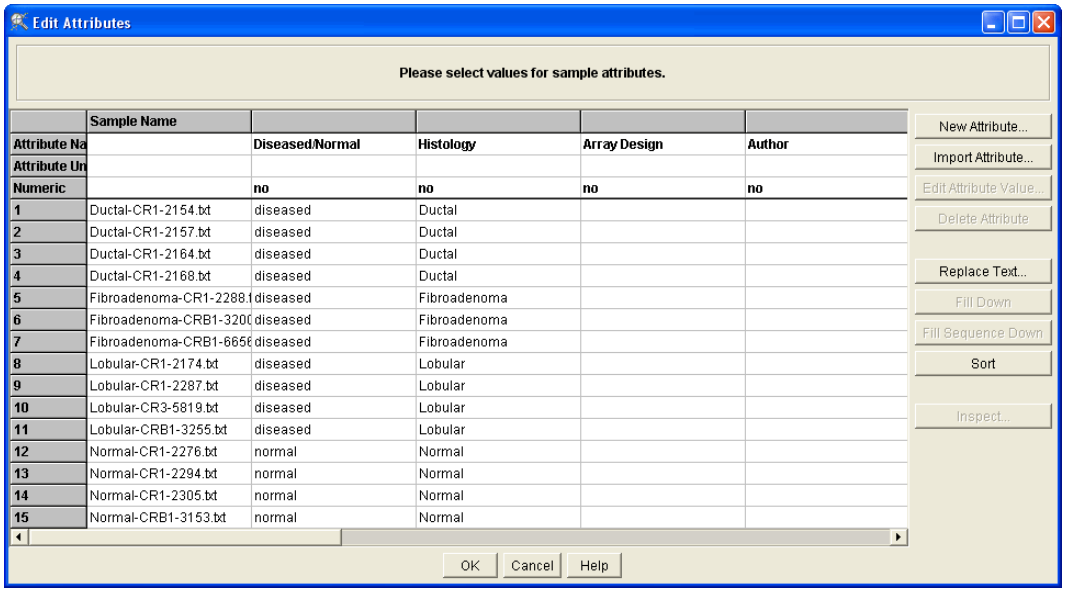
To create a new sample attribute:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.
2. Use a filter to select the samples you want.

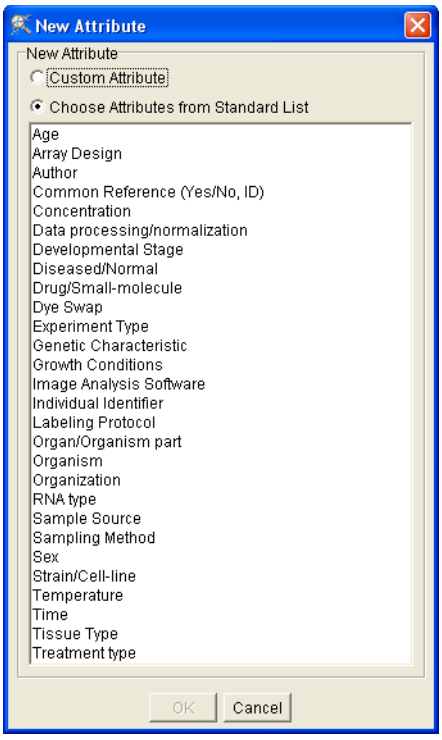
   Go to "Filtering Samples" on page 5-47 for instructions.
3. Click the **Add** button to add the samples to the Selected Samples table.
4. Click the **Edit Attributes** button.

   The Edit Attributes window opens.



5. Click the **New Attribute** button.

   The New Attribute window opens.



6. Do one of the following:
   - To add a standard attribute, select it from the menu and click the **OK** button.
   - To add a custom attribute, select the **Custom** option and click the **OK** button.

A new column appears in the Edit Attributes window.

7. Fill in the **Attribute Name** and **Attribute Units** in the new column (if applicable).
8. In the **Numeric** and **Logarithmic** rows, select **Yes** or **No** from the menus. (Click in a cell in either row to make the menu appear.) You can also paste data in the **Sample** cells.
9. Fill in values for each sample of the attribute.
10. Do one of the following:
    - To change the attributes for your currently selected sample, click the **OK** button.
    - To discard the changes, click the **Cancel** button.

Copying and Pasting Data in the Edit Attributes Window

You can paste in columns of information by clicking the cells of the Sample section. For example, if you have an Excel spreadsheet and want to copy and paste a column from it, you could copy a large section of column and paste it into the new column. You can also copy information out. You can only add columns (parameters and parameter values), you cannot add rows (samples) into this table.

If you have multiple columns you want to copy into the Edit Attributes window, you should first add as many columns as you want and then copy the contents.

You can also copy the attribute name, attribute units, and numeric fields. In the spreadsheet, create the columns you want to add. Add five lines to the top of the columns and include the attribute name, the attribute units, and the numeric state (yes/no). Paste the complete contents into the file by selecting the attribute name cell in the Edit Attributes window and then selecting the **Paste** command the menu.

Deleting Sample Attributes

To delete a sample attribute:

1. Select **Experiments** > **Sample Manager**.

   The Sample Manager window opens.
2. Use a filter to select the samples you want.

   Go to “Filtering Samples” on page 5-47 for instructions.
3. Click the **Add** button to add the samples to the Selected Samples table.
4. Click the **Edit Attributes** button.

   The Edit Attributes window opens.
5. Click the gray column bar above the attribute you want to delete.
6. Click the **Delete Attribute** button.

Using the Standard Attributes Editor

GeneSpring comes with a set of standard attributes. These attributes are available for use in any experiment in GeneSpring. You can add as many additional attributes as you like, or edit existing attributes, using the Standard Attribute Editor (Figure 5-23).

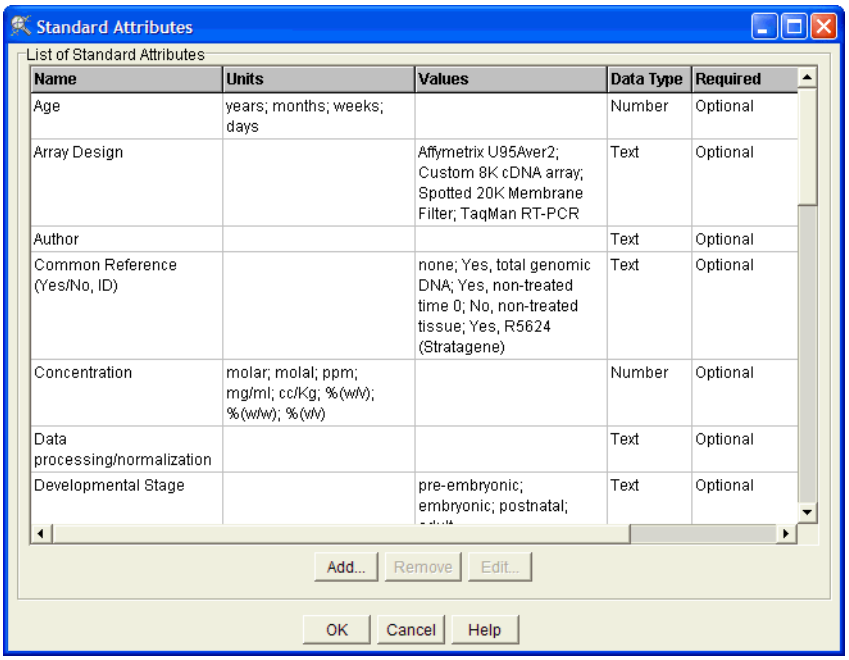


Figure 5-23 Standard Attributes Editor Window

Adding Standard Attributes

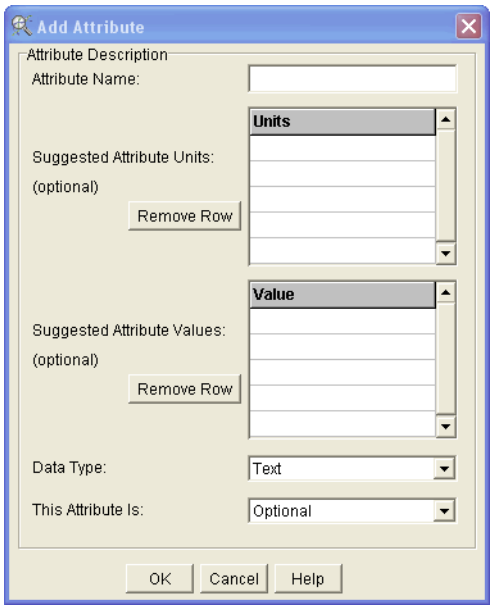
To add a standard attribute:

1. Select **Edit > Standard Attributes**.

   The Standard Attributes window opens.

2. Click the **Add** button.

   The New Attribute window opens.



3. Enter a name for the new attribute.
4. To add a suggested unit of measurement for the attribute, such as “minutes” or “ppm”, click in a row of the **Suggested Units** table and enter a unit type.

   You can add as many suggested units as you like. These units will be selectable by users when they assign this attribute to a sample.
5. To add a suggested value for the attribute, such as “Control Group” or “Martian”, click in a row of the **Suggested Values** table and enter a value.

   You can add as many suggested values as you like. These values will be selectable by users when they assign the attribute to a sample.
6. In the **Data Type** menu, select the data type (Text or Numeric)
7. In the **This Attribute Is** menu, specify whether the attribute is **Required**, **Recommended**, or **Optional**.
8. Click the **OK** button.

Editing Standard Attributes

To edit a standard attribute:

1. Select **Edit > Standard Attributes**.

   The Standard Attributes window opens.
2. Select the attribute you want to edit and click **Edit**.

   The Edit Attribute window opens.
3. Make the changes you want.
4. Click the **OK** button.

Deleting Standard Attributes

To delete a standard attribute:

1. Select **Edit > Standard Attributes**.

   The Standard Attributes window opens.

2. Select the attribute you want to delete and click the **Remove** button.

Setting Up Experiment Interpretations

Experiment Interpretations tell GeneSpring how to display your experiment parameters and handle normalized values. This section explains how to set up experiment interpretations. It covers the following topics:

- Parameters and Experiment Interpretations
- Experiment Interpretation Window
- Changing Experiment Interpretation
- Finding Experiment Interpretations
- Deleting Experiment Interpretations
- Vertical Axis Modes
- Parameter Display Settings

Parameters and Experiment Interpretations

Gene Expression analysis studies usually try to determine a group of genes that are differentially expressed when the expression values between two or more groups (conditions) are compared. The definition of those groups is done through the definition of experiment parameters. GeneSpring uses the idea of “Experiment Interpretations” to implement the separation of the samples into different groups.

A GeneSpring experiment interpretation is a way to define an experiment. A GeneSpring experiment is a group of related samples, that are separated into one or more groups. The separation into groups is defined by experiment parameters. One GeneSpring experiment can have one or more experiment interpretations. This will allow you to set up your experiment in more than one way.

For example, if you’re researching cancer, and you obtained a number of samples you can divide the biological samples in more than one way:

- Cancer Outcome: Malignant, Benign, or Normal
- Disease state: Cancerous or Normal
- Concertize: Tumor, Metastatic, or Normal

One way to analyze this data is to determine if there are any genes that are differentially expressed between cancerous and normal samples. However, another interesting research question might be to find if there are any genes that are differentially expressed between malignant or benign cancer samples.

In GeneSpring, you can perform any of these kinds of analyses by creating different interpretations for each of the different types of questions. In the example above, you would create three experiment parameters, with their respective values, and create three different experiment interpretations for each of the research questions.

It is also possible to combine two or more parameters and create unique groups (conditions) from them. For example, if you had samples from cell lines, as well as from the patient samples, you could define another experiment parameter, called “CellLine” that would define the cell line of the sample.

With this parameter defined, you could ask questions like, “Are there any genes differentially expressed between cancer and normal cells from cell lines, and are those genes different from those that are differentially expressed between cancer and normal cells from patient samples?”. Using parameters in GeneSpring will make answering these questions very easy.

For more information about setting up experiment interpretations, go to “Setting Up Experiment Interpretations” on page 5-65.

Experiment Interpretation Window

The Experiment Interpretation window (Figure 5-24) lets you determine how an experiment is to be displayed. You can change the upper and lower bounds of the vertical axis of your graph, the mode used to represent your data, whether to turn on the Cross-Gene Error Model, how you want to view each parameter, and which flagged measurements to display.

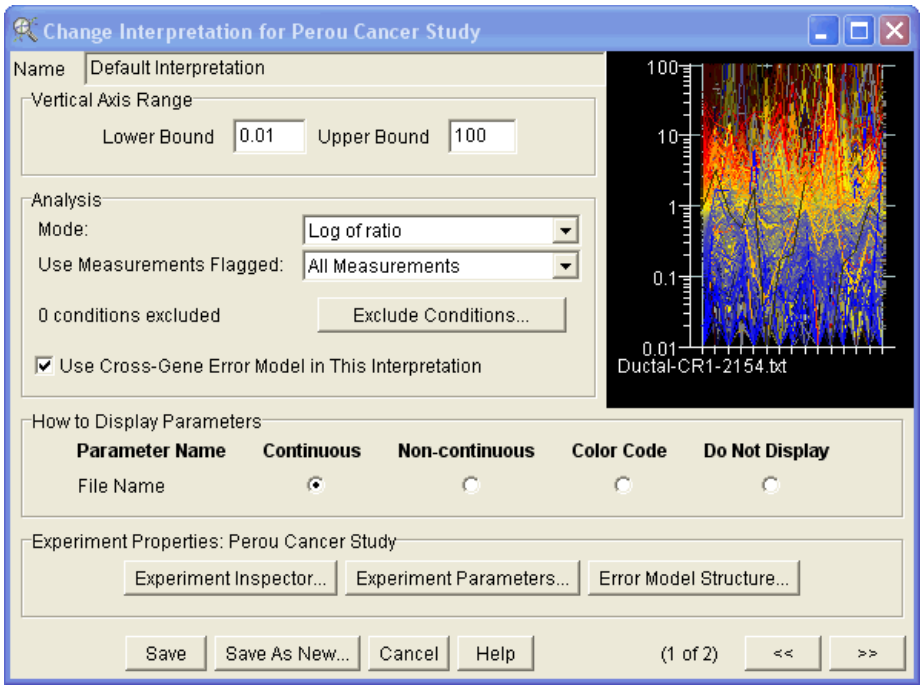


Figure 5-24 Experiment Interpretation Window

Changing an experiment interpretation is useful not only for customizing initial display settings, but also because statistical analysis techniques in GeneSpring are carried out based on how your data is characterized in the interpretation. Because of this, it can be valuable to set up more than one experiment interpretation, then perform analyses on each one to compare the results of statistical testing on data that has been grouped and characterized in different ways.

When you load your experiment, GeneSpring automatically creates two interpretations: a Default Interpretation and an All Samples interpretation. The Default Interpretation is the first item listed under the experiment in the Navigator. It may be most convenient to set up your most frequently used interpretation as your Default Interpretation. You can rename the Default Interpretation, but you cannot delete it. The All Samples interpretation makes all parameters non-continuous, so that each sample is viewed and analyzed individually. The All Samples interpretation cannot be changed, renamed or deleted.

Vertical Axis Modes

The Experiment Interpretation window lets you change experimental interpretations, including the vertical axis range and analysis mode. For example, Figure 5-25shows the Gene list “like CLN1” graphed using the [signal/control] formula. The Y-axis is graphed from 0 to 5.

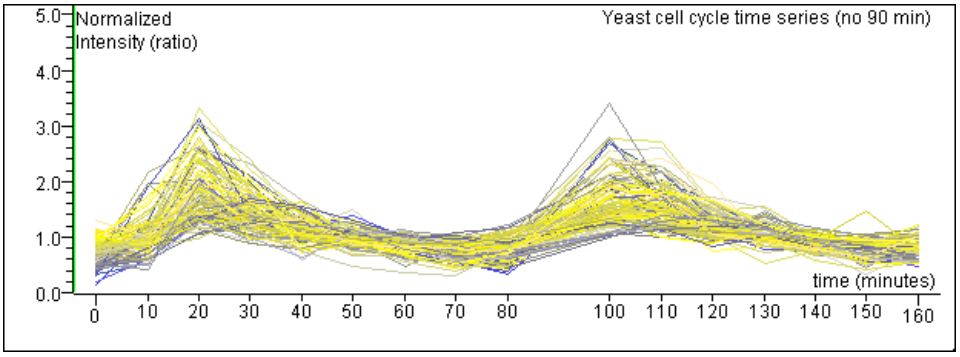


Figure 5-25 Gene List “like CLN1” Graphed with Signal/Control Formula

The “ratio” in the ratio mode refers to the way the normalized value is calculated. The normalized value is determined by dividing the signal (raw data) by the control strength.

In a one-color experiment the control strength refers to the denominator used to normalize the raw data. In a two-color experiment control strength refers to the control channel. When data is reported as the signal divided by the control, it is assumed that all expression values are positive.

The number 1 is considered normal expression; any expression value above one is overexpressed, and all underexpressed data is less than one, but greater than zero. This means that all underexpressed data appears flattened because it has to graphically fit between zero and 1, whereas overexpressed data takes up a much larger percentage of the graph (from 1 to positive infinity).

Raw signal values that are negative produce normalized values that are negative. In ratio mode, Y-axis in the graph mode is a linear scale, typically from 0 to some positive value.

For information on handling negative values, see “Affine Background Correction” on page 10-19.

Log of Ratio

The log of ratio mode (Figure 5-26) graphs normalized values (for example, the ratio of the signal to the control, not their logs), but spaces them logarithmically. The log of ratio interpretation solves the problem mentioned above under “ratio”, where all underexpressed data appears flattened because it has to graphically fit between zero and 1. In this mode underexpressed genes take up as much space visually as overexpressed genes. Logarithms of the expression ratios are used as the basis for statistical analysis.

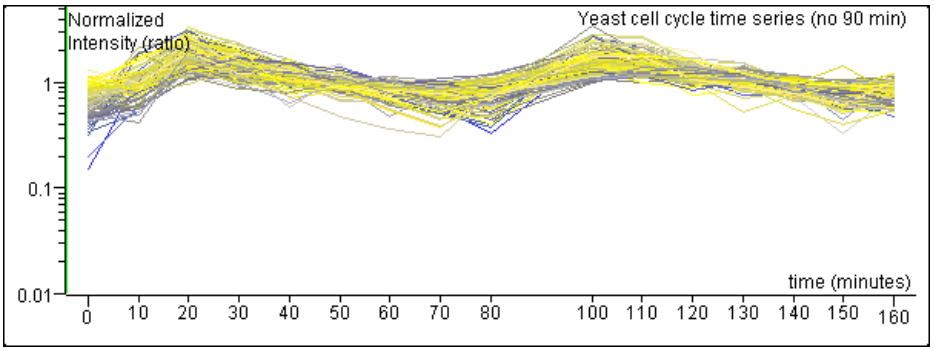


Figure 5-26 Gene list “like CLN1” Graphed with Log Ratio Formula

Note that in log of ratio mode, the lower limit of the vertical axis is 0.01. Any expression values below 0.01 are plotted as 0.01. Note also that when you export your data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01

Fold Change

Fold change mode (Figure 5-27) creates a more balanced visual representation between over- and under-expressed genes than ratio mode and emphasizes the increase and decrease of expression levels. For example, x1 would refer to normal expression, x2 to an expression level twice normal, and /2 to an expression level half normal. When using the upper or lower bound fields to change the vertical axis range enter either the ratio values in integers, or the fold change value (for example, x4 or /4). Any integers you enter are converted.

Figure 5-27 New Fold Change Image

In fold change mode, the lowest measured value is 0.01 (Table 5-2). Any values below 0.01 are calculated as 0.01. The minimum display value is /100. Note also that when you export your data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01.

Table 5-2 Fold Change Measurements

Ratio Numbers	Display
-5	/100
0	/100
.01	/100 (this is the lower cutoff)
.25	/4
.33	/3
.5	/2
1.5	x1.5
3	x3
5	x5

In addition, the values on the vertical axis are not the same as those used for subsequent analyses or those that are exported using the Copy Annotated Gene List function. In fold change mode, the values are stored as 1-N for values greater than 1 and 1-1/N for values less than 1(where N is the normalized signal). These values can be thought of as representing the distance away from normality (for example, the point that represents neither over-expression nor under-expression). These values may not be particularly useful for most users.

Parameter Display Settings

GeneSpring offers several methods for displaying a parameter: a continuous element, a non-continuous element, a replicate (or hidden) element, or a color code. When you create a new experiment, the default interpretation will use the first parameter as a continuous element. All other parameters will be set to replicate.

Continuous Elements

In a continuous variable, each parameter value exists in series on a continuum with the other values in that parameter, rather than as discrete points. Each value is related to the values on either side of it and adjacent data points are connected together by lines. Typically, continuous variables are numeric. This requires that the values be in a particular order. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order.

When graphing by a continuous parameter each value is placed on the X-axis in order from left to right. You can change this default order. See “Changing the Order of Parameters on the Horizontal Axis” on page 5-44 for more details.

Non-Continuous Elements

In a non-continuous (or set) variable, each parameter value exists independent of the others, as a discrete point. When a non-continuous element is graphed, each parameter value is placed on the horizontal axis, in order from left to right. GeneSpring automatically orders numerical parameters from highest to lowest. Non-numerical parameters are in alphabetical order. See “Changing the Order of Parameters on the Horizontal Axis” on page 5-44 if you need non-numerical parameter values to be graphed in a particular non-alphabetical order.

When displaying data from a non-continuous parameter, data points are graphed in histograms, as discrete points. A gene deletion is a simple example of a non-continuous element, but it is by no means the only possible non-continuous parameter. A non-continuous parameter is occasionally referred to as a set when there are other parameter display options employed (especially when a continuous parameter is used) because the non-continuous parameter separates the data into a series of discrete graphs viewed next to each other on the same window. When a continuous parameter is used in conjunction with a non-continuous parameter each discrete graph contains all of the values of the continuous parameter, making each of the separate graphs look like a set of parameter values.

Hidden Elements

Parameters defined as replicates are averaged together and appear as a single parameter. A parameter defined as a replicate is graphically a hidden variable. Defining a parameter as a replicate is the easiest way to deal with repeated samples inside GeneSpring.

The equation used for averaging repeated samples is exactly the same one used to average repeated measurements in a raw data file. See “Handling Repeated Measurements” on page 10-26 for more information. The only difference is that averaging of repeated parameters is done after the raw data has been normalized.

Color Code

The Color Code display setting colors genes by parameter. A color code is used for experimental parameters in which parameter values exist independently of one another, but are not unrelated. When data in the Genome Browser is colored by parameter, GeneSpring orders the parameter values from top to bottom in the Colorbar. See “Coloring by Parameter” on page 8-37 for details. Values are listed in alphabetic or numerical order.

Each color represents a category (or set of categories). When coloring the browser display by parameter, each value defined as a condition is assigned a color and every data point described by that parameter is drawn in that parameter’s color. This can be referred to as **Color by Parameter.** This display option shows the same gene multiple times. The number of times a single gene is drawn is equal to the number of values defined as conditions.

When the Genome Browser display is colored using a color option other than Color by Parameter, it is impossible to visually distinguish which value a particular gene line or gene point represents, although separate gene lines for each value defined as a condition are still drawn. See “Changing the Order of Parameters on the Horizontal Axis” on page 5-44 for details on how to change that order.

Individual patients, or strain types, are variables commonly defined as color codes (conditions) because, although they are different values, it is interesting to see them visually compared to one another. It is likely the expression patterns of individual patients with the same disease will react in a similar way under similar conditions. Often it is when the expression patterns are not similar that the results are interesting. This is where graphs of parameter-values defined as color-coded conditions are useful as they allow you to easily compare varying conditions of the same gene.

Opening the Experiment Interpretation Window

To open the Experiment Interpretation window:

- Do one of the following:
  - Select **Experiments > Experiment Interpretation**.
  - In the **Navigator**, open the **Experiments** folder, right-click the interpretation you want, and then select the **Inspect** command.
  - In the **Navigator**, open the **Experiments** folder, locate the interpretation you want, and then double-click the interpretation.

Changing Experiment Interpretation

To change experiment interpretation:

1. Open the **Experiment Interpretation** window.

   Go to "Opening the Experiment Interpretation Window" on page 5-71 for instructions.

2. From the **Mode** menu, choose a data display mode for the vertical axis. You have the following choices:

   - Ratio (signal/control)
   - Log of ratio
   - Fold change

   The mode you choose is used in such statistical procedures as Statistical Group Comparison, k-means Clustering, Self-organizing Maps, and Principal Components Analysis. See below for details on these modes. Choose the lower and upper bounds of the vertical axis in the fields provided. For more information about modes, go to "Vertical Axis Modes" on page 5-67.

   These lower and upper bounds are used in the Graph views and will limit the boundaries of the vertical axis. These values are the default values for the interpretation, and they can be overridden manually by setting the display options in the Graph views.

   For more information about Graph views, go to "Using the Graph View" on page 8-9. For more information about the vertical axis, go to "Vertical Axis Modes" on page 5-67.

3. Depending on your instrumentation, you may have flags indicating the degree to which your data is reliable. If you have flags, choose from the **Use Measurements Flagged** menu to limit data based on these flags.

4. (Optional) To use the Cross-Gene Error Model, check the **Use Cross-Gene Error Model** box.

   If the number of replicates in your condition is low (lower than 3), the Cross Gene Error model might give you some opportunity to obtain some more confidence values for the expression values. By default the error model is used, so make sure it is appropriately setup. Go to "Using Cross-Gene Error Models" on page 5-73 for more information.

5. Choose a mode for each parameter: **Continuous Element**, **Non-continuous**, **Color Code,** or **Do Not Display**. Note that if you choose Color Code, you must also select **Colorbar > Color by Parameter**. For more information on parameters, go to "Parameter Display Settings" on page 5-70.

6. To make any additional changes in your experiment before you continue, click any of the buttons in the Experiment Properties panel. Your options are:
   - **Experiment Inspector**—Opens the Experiment Inspector window. For a detailed description of this window, see "Using the Experiment Inspector" on page 7-25.
   - **Experiment Parameters**—Opens the Experiment Parameters window. For a detailed description of this window, see "Using the Experiment Parameters Window" on page 5-39.
   - **Error Model Structure**—Opens the Cross-Gene Error Model window. For a detailed description of this window, see "Using Cross-Gene Error Models" on page 5-73.
7. When you are done, name your interpretation.
8. Do one of the following:
   - To update your current interpretation, click the **Save** button.
   - To create a new interpretation, click the **Save As** button.

Finding Experiment Interpretations

To find an experiment interpretation:

- Do one of the following:
  - Select **Experiments** > **Experiment Interpretation**.
  - In the **Navigator**, open the **Experiments** folder, right-click the interpretation you want, and then select the **Inspect** command.
  - In the **Navigator**, open the **Experiments** folder, locate the interpretation you want, and then double-click the interpretation.

Deleting Experiment Interpretations

To delete an experiment interpretation:

- In the **Navigator**, open the Experiments folder, right-click the interpretation you want, and then select the **Delete** command.

Using Cross-Gene Error Models

The ability to estimate measurement and sample-to-sample variation in microarray-based experiments is often compromised by the fact that the cost (in both time and materials) of performing large numbers of replicate samples is quite high.

If the Cross-Gene Error Model (Figure 5-28) function is enabled, GeneSpring accounts for error instead by assuming that the amount of variability is a function of the control strength within all the measurements for a single experimental condition. The advantage of making this assumption is that the number of measurements used to estimate the global error is equal to the total number of genes on any given chip.

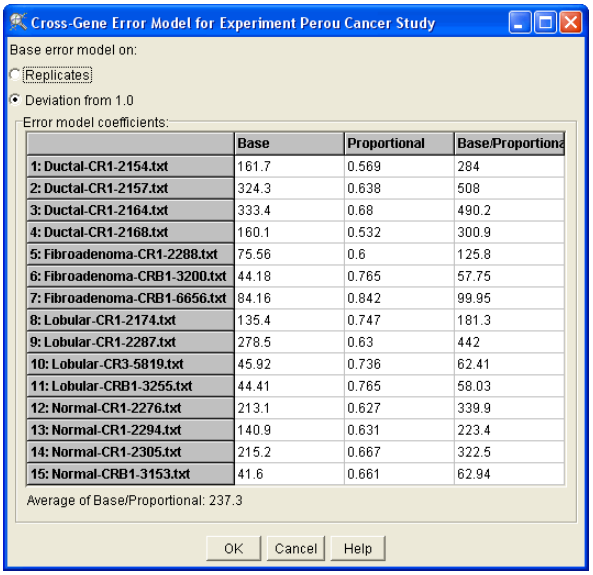


Figure 5-28 Cross-Gene Error Model Window: Deviation from 1.0

If you have replicates for each condition of your experiment, you can use the Replicates option (Figure 5-29) to specify the parameters that differentiate the groups of replicates.

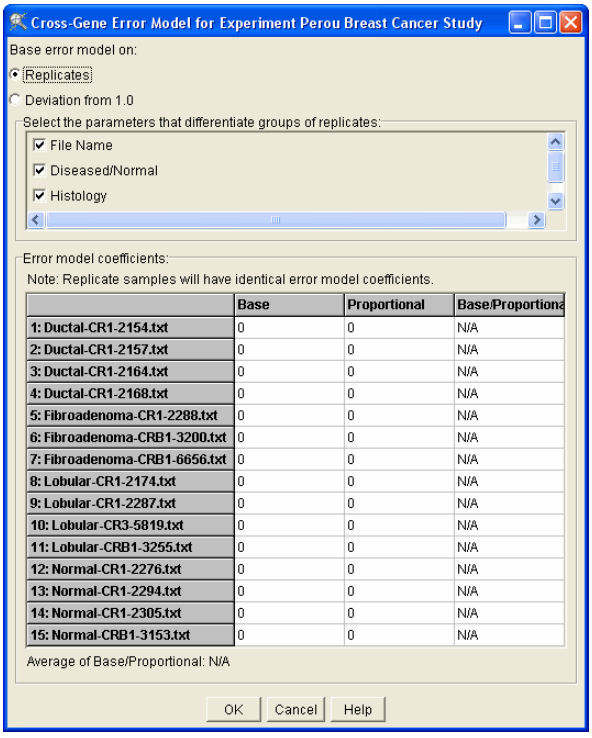


Figure 5-29 Cross-Gene Error Model Window: Replicates

In addition, measurement precision information supplied by the scanner software or independently by the user can be loaded into GeneSpring via the “Signal Precision” column type in the Column Editor. The value given in this column is interpreted as the standard deviation of the raw measured value.

The sample-to-sample variability includes the effect of both types of variation, and the statistical separation of these effects is called variance components analysis. The GeneSpring Cross-Gene Error Model performs this variance components analysis, and uses the estimates of these two components of variation to accurately estimate standard errors and compare mean expression levels between experimental conditions.

Separate estimates of two different kinds of random variation are used to estimate the variability in gene expression measurements:

- **Measurement variation—**This comprises the lowest level of variation, corresponding to the variation of the measurement of a gene on a single chip around the true value that would be achieved by a perfect measurement of the expression level of the gene for that sample.
- **Sample-to-sample variation—**This is the variation between samples in the same condition. This represents biological or sampling variability, such as variability between multiple subjects in a condition, between multiple physical samples for an experimental subject or patient, or between multiple hybridizations of a physical sample. GeneSpring can represent any one of these kinds of variability, depending on the types of replicate samples you have specified in your interpretation and in the error model dialog. GeneSpring assumes all replicate samples in the same condition correspond to one kind of variability.

When you enable the Cross-Gene Error Model it is used as the basis for:

- Standard deviation, representing the variability of individual population members.
- Standard error, representing the precision of the mean of the gene expression measurements in the condition with respect to the true condition mean.
- Error bars corresponding to standard deviation or standard error in the Graph view and Gene Inspector.
- T-test p-value, representing the statistical test of differential expression for a specific condition.
- Color by significance, coloring according to the t-value from the t-test of differential expression.
- Finding differentially expressed genes using the Statistical Group Comparison, if the error model option is chosen

Defining the Cross-Gene Error Model

To define the Cross-Gene Error Model:

- Do one of the following:
  - Select **Experiments > Cross-Gene Error Model**.

    The Cross-Gene Error Model window opens.
  - Select **Experiments > Experiment Interpretations** and click **Error Model Structure** in the Experiment Interpretations window open.

    The Cross-Gene Error Model window opens.

Enabling the Cross-Gene Error Model

To enable the Cross-Gene Error Model:

1. Open the **Cross-Gene Error Model**.

   Go to "Defining the Cross-Gene Error Model" on page 5-76 for instructions.
2. If you have replicates for each condition, do the following:
   a. Select the **Replicates** option.
   b. Select the parameters that define the grouping.

      GeneSpring considers each sample with the same parameter value as a replicate.
   c. Click the **OK** button.
3. If you do not have replicates for each condition, select the **Deviation from 1.0** option and click the **OK** button.

   Note: Double-click on a row to view that sample in the Sample Inspector.
4. Select **Experiments > Experiment Interpretation**.

   The Change Interpretation window opens.
5. Select the **Use Cross-Gene Error Model** option.
6. Do one of the following:
   - To update your current interpretation, click the **Save** button.
   - To create a new interpretation, click the **Save As** button.

Technical Details for the Cross-Gene Error Model

The two-component model for estimating variation from control strength is known as the Rocke-Lorenzato model. The two components are an absolute error component that dominates at low measurement levels, and a relative error component that dominates at high measurement levels. The formula for the error model for raw (pre-normalization) expression levels can be written as:

$$\sigma_{RAW} = \sqrt{a^2 + b^2 S^2}$$

where σ_{RAW} is the measurement standard error of the raw expression data, S is the measurement level (control strength), and a and b are the fitted coefficients of the model.

Expressed in terms of the normalized expression levels, which are the result of dividing raw expression levels by control strength, the standard errors can be written as:

$$\sigma_{NORM} = \sqrt{\frac{a^2}{s^2} + b^2}$$

Before fitting the error model, the genes are ordered by their control strengths. A median variance and median control strength is calculated for each non-overlapping set of eleven genes. If replicates are used, this variance is the standard error of the samples in the current condition. If the “deviation from 1” option is selected, error is approximated by using the median deviation from 1.0.

The goal in this step is to remove outliers (when replicates are being used) and to disregard genes whose high or low expression level is the result of biological activity. In the absence of replicates the working assumption is that the vast majority of the genes do not change over the conditions in the experiment, and thus deviation from one represents error in a gene whose expression level changes little over the course of the experiment. Then an iteratively reweighted linear regression of variation or squared deviation versus squared control strength is fitted to estimate the parameters.

Estimation of the 2-level variance components model is done by the method of moments. In order to eliminate negative estimates of variance components, within-sample variation is taken as a lower bound on total between-sample variation. Different sources of information in the analysis are weighted by their appropriate statistical degrees of freedom. Precision estimates based on replicate genes or samples are assigned degrees of freedom equal to the number of replicates minus one. User-supplied precision values, if available, are assigned 1 degree of freedom.

Cross-Gene Error Models, if used, are assigned an equal number of degrees of freedom as the direct variability estimates for that gene. Between-sample analyses are done according to the interpretation mode (ratio, log, fold). Within-sample variability is calculated in terms of normalized ratio expression, and translated as necessary to the interpretation mode by use of the delta method.

Results of the variance components analysis are used to estimate standard deviations and standard errors, according to the grouping of samples into conditions as specified by the experiment interpretation. Two different types of interpretation affect the assumed context of the calculation:

- **Single-sample interpretation—**If all conditions contain only one sample (for instance the. “All Samples” interpretation), precision calculations are based solely on the estimated within-sample measurement variation. The error bars, standard deviations, and standard errors represent the variability of all possible measurements on this specific sample.
- **Multi-sample interpretation—**If at least one condition contains multiple samples, precision calculations for all samples are based on the combined within-sample and between-sample variation, and error bars, standard deviations, and standard errors represent the variation of measurements of samples representing the population of all possible samples in the condition.

  In a multi-sample interpretation, if no replicate samples are available for a specific condition, then no error calculations are made and no error bars are shown, since there is no information available on the variability of that condition.

References

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, John Wiley and Sons, New York.

Milliken, G. A. and Johnson D, E. (1984) *Analysis of Messy Data*, Volume 1: Designed Experiments. Wadsworth, Inc. Belmont, California.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2:110-14.

Tuimala, Jarno and Laine, M. Minna (Eds). (2003) *DNA Microarray Data Analysis*, CSC - Scientific Computing Limited. Picaset Oy, Helsinki.

6

Managing Genomes

This chapter explains how to manage information about genomes in GeneSpring. It covers the following topics:

- Overview
- Managing Genomes
- Inspecting Genomes
- Managing Genes and Annotations

Overview

GeneSpring provides a number of tools for organizing, managing, navigating, and editing genomes. These tools include:

- **Genome Manager**—Lets you reorganize the genome hierarchy, rename local genomes, delete genomes, and export genomes as GeneSpring zip files.
- **Genome Inspector**—Lets you view a summary of the contents in the Master table of genes, a history of changes to the genome, the GenBank/EMBL files associated with a genome, the full nucleotide sequence (if available), web links, and homology tables associated with the genome.
- **Master Table of Genes Editor**—Contains all the annotations associated with genes in a given genome. Using tools in the Master Table of Genes Editor, you can add genes, delete genes, create annotations, edit annotations, add non-genomic elements, and search for genes without the need to manually edit the files outside of GeneSpring.

This chapter provides information about using these tools.

Managing Genomes

This section explains how to use the Genome Manager to manage genomes in GeneSpring. It covers the following topics:

- Genome Manager Window
- Opening the Genome Manager
- Opening Genomes
- Importing GeneSpring Zip Files
- Exporting GeneSpring Zip Files

- Adding New Folders
- Deleting Folders
- Renaming Folders
- Renaming Genomes

The Genome Manager also provides tools for managing data in Signet. For complete information on working with data in Signet, go to Chapter 18, “Using Signet”.

Genome Manager Window

The Genome Manager (Figure 6-1) provides tools for reorganizing the genome hierarchy, renaming local genomes, deleting genomes, and exporting genomes as GeneSpring zip files.

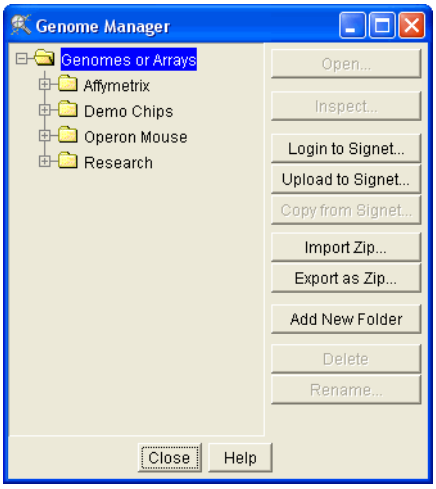


Figure 6-1 Genome Manager Window

Navigator Panel

The Navigator, located on the left side of the window, shows the genomes or arrays in GeneSpring.

Buttons

- **Open**—Opens a Genome Browser window for the selected genome. If the selected genome is already open, it opens the associated Genome Browser to the front. This button is only enabled when one, and only one, genome is selected.
- **Inspect**—Opens the Genome Inspector for the selected genome. This button is only enabled when only one genome is selected.

  Go to “Inspecting Genomes” on page 6-10 for more information.
- **Login to Signet**—Opens the Login to Signet window enabling you to log in to Signet.

  Go to “Logging in to Signet” on page 18-4 in Chapter 6, “Managing Genomes”for instructions.

- **Upload to Signet**—Enables data objects in a genome to be uploaded to Signet. The folder hierarchy of the genomes is lost after the folder is uploaded. This function requires access to Signet.

  Go to "Uploading Data to Signet" on page 18-5 in Chapter 6, "Managing Genomes"for instructions.

- **Copy from Signet**—Makes a copy of a genome if it does not exist locally. If a local copy already exists, this option makes a new copy of the genome.

  Go to "Changing Ownership of Data in Signet" on page 18-10 in Chapter 6, "Managing Genomes"for instructions.

- **Import Zip**—Opens the Import Zip file window which allows you to import the genome zip file.
- **Export as Zip**—Saves a zip of the selected genome but not the contents of the genome). This button is enabled when you select a single genome or a folder of genomes.
- **Delete**—Deletes the selected file or folder. This button is enabled when only one genome or one folder is selected.
- **Add New Folder**—Adds a new folder within the selected folder. This button is enabled when one folder is selected.
- **Rename**—Renames the selected genome. This button is enabled if one local genome is selected.

Menu

Right-clicking on a file or folder in the Genome Manager displays commands for performing different actions. Table 6-1 describes the commands that can appear in this menu.

Table 6-1 Genome Manager: Menu

Command	Description
Inspect	Opens the Genome Inspector.
Upload to Signet	Uploads data from GeneSpring to Signet.
Copy from Signet	Copies a genome from Signet to GeneSpring.
Assign Project	Adds a project to an experiment.
Rename	Renames a folder or file.
Delete	Deletes the selected file or folder.
Export as Zip	Exports the selected data as a zip file.
Import Zip (on folders only)	Imports the selected zip file.
Add New Folder (on folders only)	Creates a new folder.

Drag-and-Drop Actions

The following drag-and-drop actions are supported in the Genome Manager:

- Drag and drop a genome into a folder.
- Drag and drop a local genome into a Signet folder. This action displays the Bulk Upload window for that genome.
- Drag and drop a Signet genome into a local genome folder. This action copies the genome from Signet to the designated location.
- Drag and drop a GeneSpring genome zip file from your desktop (or other location) into the Genome Manager. This action adds the genome to GeneSpring.

  Note: You cannot drag a genome from the Genome Manager to the desktop Use the Export Zip file command to perform this action instead.

Opening the Genome Manager

To open the Genome Manager:

- Select **File** > **Genome Manager**.

Opening Genomes

This section explains how to open genomes that reside in GeneSpring or Signet.

To open a genome:

1. Select **File** > **Genome Manager**.

   The Genome Manager window opens.

2. In the **Genomes or Arrays** folder, select the genome you want.
3. Click the **Open** button.
   - If a window opens and informs you that the genome is loading, you have successfully opened the genome. After the genome has loaded, a new Genome Browser opens and displays your data.
   - If an error message opens, you need to resolve the problem before you can proceed. Go to "Signet Warnings" on page 6-4 for more information.

Signet Warnings

If you are logged into Signet and attempt to open a genome or array, GeneSpring verifies the date and time stamps of the file. If there is a mismatch, GeneSpring displays a warning message, as shown in Table 6-2. If you receive a mismatched gene message while attempting to open a genome from Signet, you'll need to resolve the problem before proceeding.

Table 6-2 Opening Genomes from Signet

Time Stamps	Gene List Comparison	Result
Local is more recent	Same list of genes	No messages
Signet is more recent	Same list of genes	Update Genome message
Local is more recent	Mismatched list of genes	Mismatched genes message
Signet is more recent	Mismatched list of genes	• Update Genome • Mismatched genes message (only appears if the update doesn't fix the problem, or if you chose not to update)
Local and Signet have the same timestamps	Same list of genes	No messages
Local and Signet have the same timestamps	Mismatched list of genes	Mismatched genes message

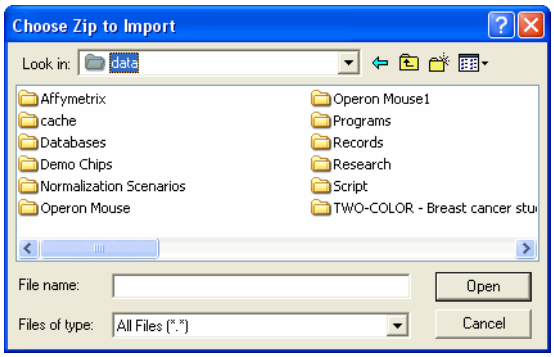
Importing GeneSpring Zip Files

GeneSpring zip files are specially designed compressed files that are automatically recognized by GeneSpring. GeneSpring zip files can be created by GeneSpring and Silicon Genetics. GeneSpring zip files can be used to update and add new functionality to GeneSpring.

To import a GeneSpring zip file:

1. Do one of the following:
   - Select **File > Genome Manager** and click the **Import Zip** button.
   - Select **File > Import GeneSpring Zip**.

     The Choose Zip to Import window opens.



2. In the **Look In** list, select the directory you want.
3. Navigate to the folder you want.
4. Select the file you want to import.

5. Click the **Open** button.

   The Importing Zipped Genome progress bar opens.

Exporting GeneSpring Zip Files

GeneSpring lets you compress one or more genomes and export them to a directory of your choice. Once a zip file is created, you can send it to colleagues via email or other methods. Users can then open the zip files directly in GeneSpring and install the new genomes.

To export a GeneSpring zip file:

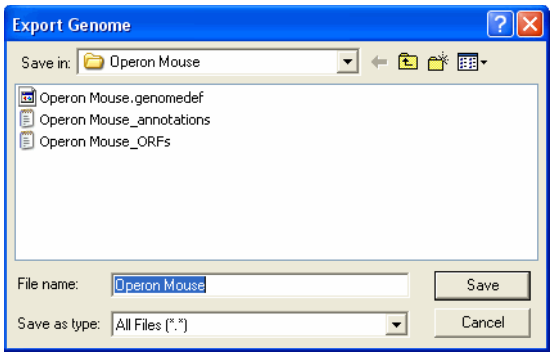
1. Select **File > Genome Manager**.

   The Genome Manager window opens.

2. In the **Genomes or Arrays** folder, select the genome you want.

3. Click the **Export as Zip** button.

   The Export Genome window opens.



4. In the **Save In** list, select the directory where you want to save your zip file.

5. Click the **Save** button.

Adding New Folders

The Genome Manager lets you add new folders to the Genomes and Arrays directory. You can use these folders to store genomes or zip files.

To add a new folder:

1. Select **File > Genome Manager**.

   The Genome Manager window opens.

2. In the **Genomes or Arrays** folder, do one of the following:

   - Select a folder, click the **Add New Folder** button, and enter the folder name.
   - Right-click a folder, select the **Add New Folder** command, and enter the folder name.

Deleting Folders

This section explains how to delete folders using the Genome Manager. Deleting a folder that contains genomes also deletes all genomes in the folder. If a folder contains genomes, two warning dialog boxes appear and require you to confirm the deletion. Go to "Deleting Genomes" on page 6-9 for more information.

To delete a folder:

1. Select **File > Genome Manager**.

   The Genome Manager window opens.

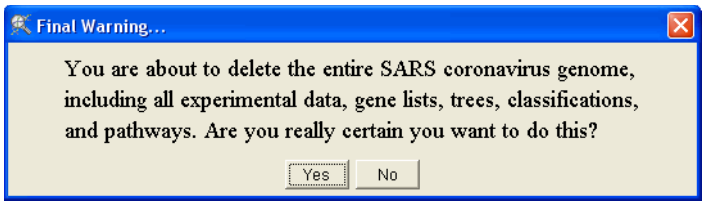
2. In the **Genomes or Arrays** folder, do one of the following:

   - Select the folder you want and click the **Delete** button.
   - Right-click the folder you want and select the **Delete** command.

     If the folder contains genomes, a warning message appears.

3. Select the **Yes** button to confirm.

   A final warning window opens.


Final Warning...
You are about to delete the entire SARS coronavirus genome, including all experimental data, gene lists, trees, classifications, and pathways. Are you really certain you want to do this?
Yes
No

4. Select the **Yes** button to confirm.

   The folder and its contents are deleted from GeneSpring.

Renaming Folders

To rename a folder:

1. Select **File > Genome Manager**.

   The Genome Manager window opens.

2. In the **Genomes or Arrays** folder, do one of the following:

   - Select the folder you want, click the **Rename** button, and enter the new name.
   - Right-click the folder you want, select the **Rename** command, and enter the new name.

Renaming Genomes

The name of a genome is of vital importance in GeneSpring. The genome name is used as an identifier to determine which experiments and other data objects are related to the genome and array.

This procedure explains how to rename a genome. You may decide to rename a genome because you have different genomes with the same name, or the current name of the genome is no longer appropriate.

Restrictions

The following naming restrictions apply:

- The new name cannot duplicate the name of a pre-existing local genome
- The new name cannot be longer then 80 characters (because of Signet)
- The new name cannot be blank

The following characters are allowed:

```
a-z, A-Z, 0-9, space, _().-& (no comma)
```

The following characters are not allowed:

```
~`!@#$%^*+={}|[]\;':"<>?,/
```

Limitations

Renaming a genome may cause the following problems (the user will receive a warning before they can modify the genome name):

- Scripts that are executing remotely will not automatically recognize the genome to which the results belong.
- Objects exported as zip files will not automatically be recognized as originating from the renamed genome.
- Bookmarks saved outside of the data directory will break.
- Signet homology tables that point to the local genome will break.
- Local and Signet genomes that are currently recognized as being the same will not be recognized as such, therefore it will no longer be possible to view your local and remote data together.

Renaming Genomes

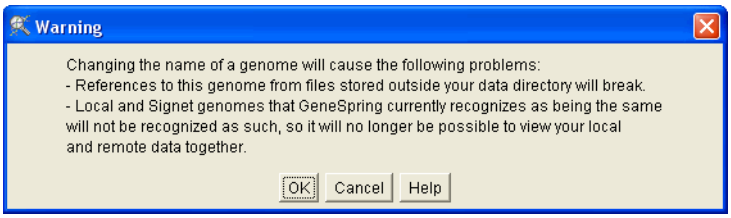
To rename a genome:

1. Open the **Genome Manager**.

   Go to "Opening the Genome Manager" on page 6-4 for instructions.

2. In the **Genomes or Arrays** folder, do one of the following:
   - Select the genome you want and click the **Rename** button.
   - Right-click the genome and select the **Rename** command.

A warning window opens.


Warning
Changing the name of a genome will cause the following problems:
- References to this genome from files stored outside your data directory will break.
- Local and Signet genomes that GeneSpring currently recognizes as being the same will not be recognized as such, so it will no longer be possible to view your local and remote data together.
OK
Cancel
Help

3. Review the warning and click the **OK** button to continue.
4. Enter the new name for the genome.

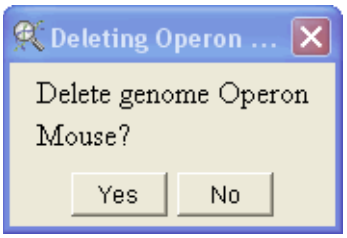
Deleting Genomes

To delete a genome:

1. Open the **Genome Manager**.

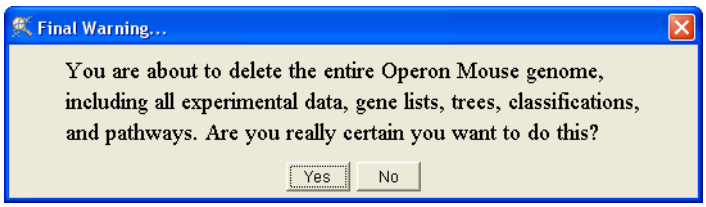
   Go to "Opening the Genome Manager" on page 6-4 for instructions.
2. In the **Genomes or Arrays** folder, do one of the following.
   - Select the genome you want and click the **Delete** button.
   - Right-click the genome and select the **Delete** command.

   A confirmation window opens.



3. Select the **Yes** button to confirm.

   A final warning window opens.


Final Warning...
You are about to delete the entire Operon Mouse genome, including all experimental data, gene lists, trees, classifications, and pathways. Are you really certain you want to do this?
Yes
No

4. Select the **Yes** button to confirm.

   The genome is deleted from GeneSpring.

Inspecting Genomes

This section explains how to use the Genome Inspector in GeneSpring. It covers the following topics:

- Genome Inspector Window
- Opening the Genome Inspector
- Editing Gene Properties
- Viewing Genome History
- Importing Genomic Sequences
- Building Homology Tables
- Managing Experiment Links

Genome Inspector Window

The Genome Inspector window (Figure 6-2) lets you:

- View a summary of the contents in the Master Table of Genes
- View a history of changes made to the genome
- View the GenBank/EMBL files associated with the genome
- Import the full nucleotide sequence (if available)
- Create, edit, or delete gene web links
- View, add, edit, and delete homology tables associated with the genome

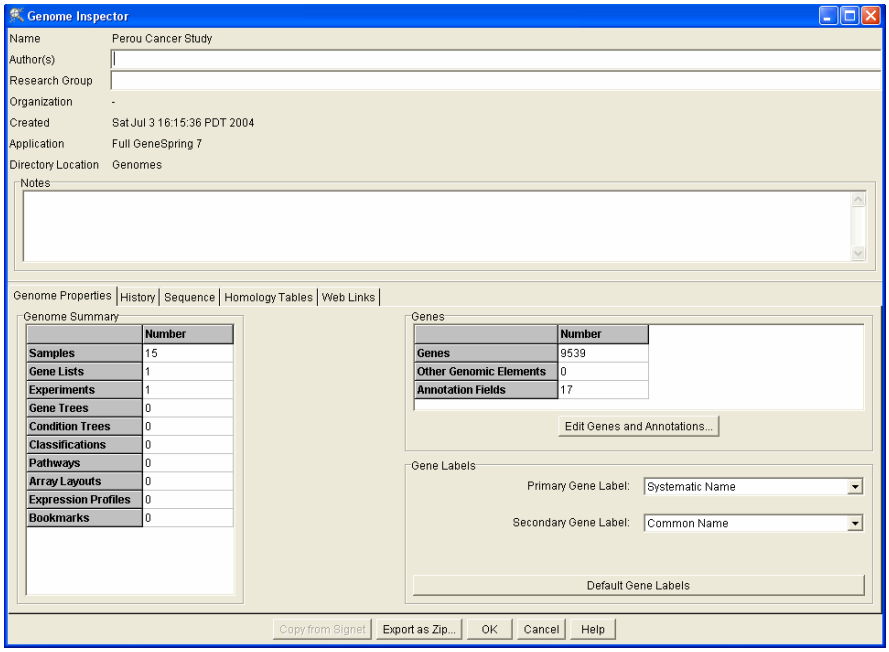


Figure 6-2 Genome Inspector Window: Sequence Tab

If you imported a file from GenBank or EMBL, the Genome Inspector window also includes the GenBank/EMBL tab (Figure 6-3).

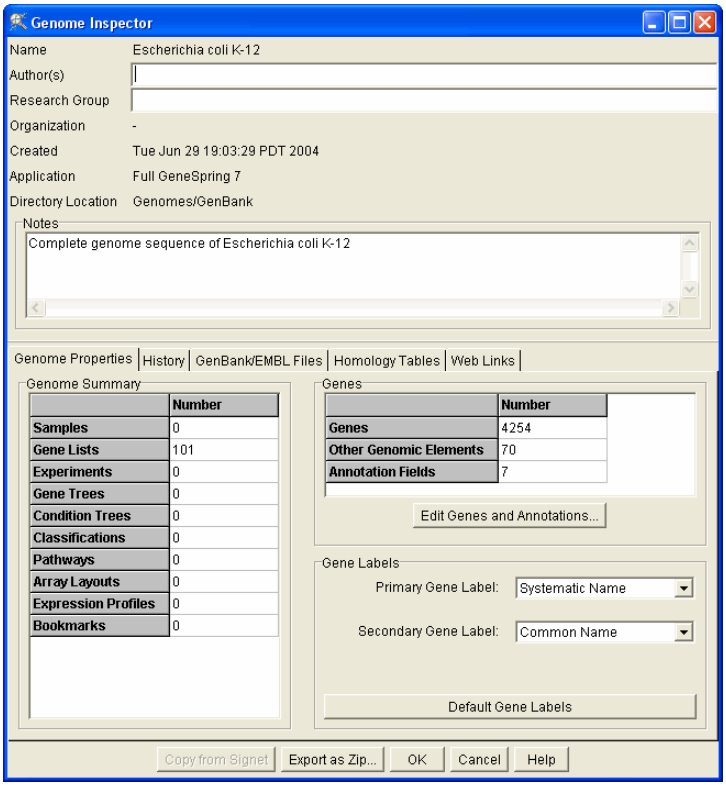


Figure 6-3 Genome Inspector Window: GenBank/EMBL Tab

The following section describes the elements in the Genome Inspector window.

Summary Information

At the top of the Genome Inspector window are fields that let you add information about the genome.

- **Name**—Shows the name of the genome.
- **Author(s)**—Lets you enter the name of the authors associated with the genome.
- **Research Group**—Lets you enter the name of the research group associated with the genome.
- **Organization**—Shows the name of the organization associated with the genome. This information cannot be edited; it is derived from the license key.
- **Created**—Shows the date and time at which the genome was imported into GeneSpring.
- **Application**—Shows the GeneSpring name and version number used to import the genome.
- **Location**—Shows the name of the directory in which the genome is saved.

- **Notes**—Lets you enter information about the genome.

Buttons

- **Copy from Signet Button**—Copies a genome from Signet and imports it into GeneSpring. This button is disabled if the genome is not on Signet, or if you are not logged in to Signet.
- **Export as Zip Button**—Exports the genome as a zip file. It does not contain the content of the genome.
- **OK**—Confirms your selection and closes the window.
- **Cancel**—Cancels your selection and closes the window.
- **Help**—Displays online Help.

Genome Properties Tab

The Genome Properties tab (Figure 6-4) provides descriptive information about the genome.

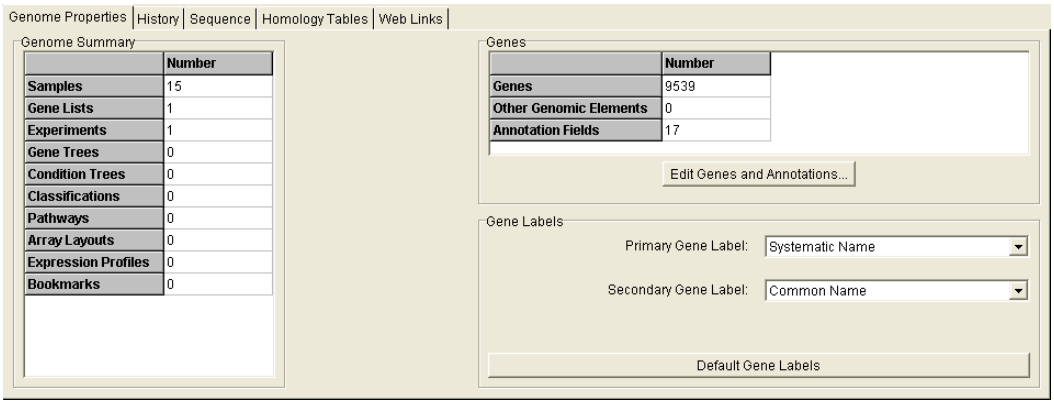


Figure 6-4 Genome Inspector Window: Genome Properties Tab

Genome Summary Table

The Genome Summary table summarizes the contents of the genome at the time the Genome Inspector was opened. This information can change over time. The table contains the following entries:

- **Samples**—Displays the number of samples associated with this genome.
- **Gene Lists**—Displays the number of gene lists associated with this genome.
- **Experiments**—Displays the number of experiments associated with this genome
- **Gene Trees**—Displays the number of gene trees associated with this genome.
- **Condition Trees**—Displays the number of condition trees associated with this genome.
- **Classifications**—Displays the number of classifications associated with this genome.
- **Pathways**—Displays the number of pathways that have been created for this genome.

- **Array Layouts**—Displays the number of array layouts associated with this genome.
- **Expression Profiles**—Displays the number of expression profiles associated with this genome.
- **Bookmarks**—Displays the number of bookmarks associated with this genome.

Genes Table

The Genes table provides a summary of the contents of the Master Table of Genes. It contains three categories:

- **Number of Genes**—Displays the number of genes associated with this genome.
- **Number of Other Genomic Elements**—Displays the number of genomic elements associated with this genomes that are not genes, like telomers or ribosomal RNA genes.
- **Number of Annotation Fields**—Displays the number of annotation fields associated with this genome.
- **Edit Genes and Annotations Button**—Opens the Master Table of Genes window.

Gene Labels Box

The Gene Labels box lets you set your own gene labels. It contains the following elements:

- **Primary Gene Label**—Sets the primary name GeneSpring displays for a gene. The default name is Systematic Name.
- **Secondary Gene Label**—Sets the secondary name GeneSpring displays for a gene. This name usually appears in parentheses. The default name is Common Name.
- **Default Gene Name Button**—Displays the settings used in the Primary Gene Label and the Secondary Gene Label lists.

History Tab

The History tab (Figure 6-5) tracks all changes made to the genome.

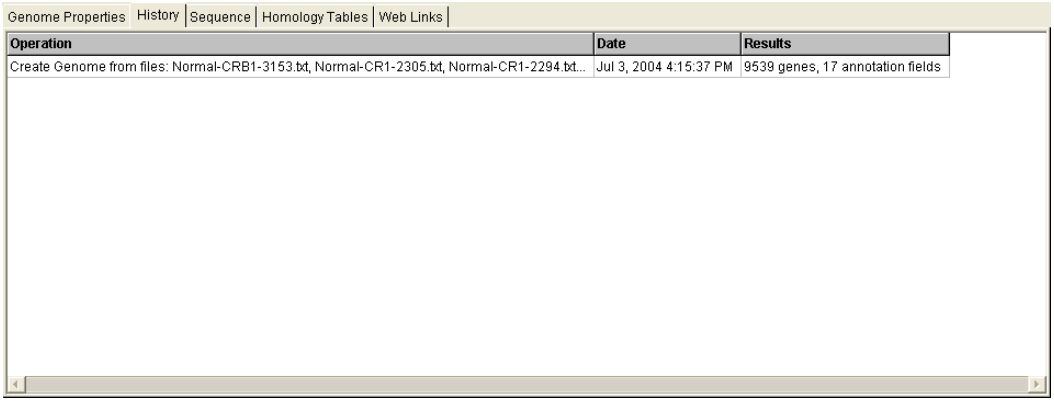


Figure 6-5 Genome Inspector Window: History Tab

The changes that can appear in the History tab include the following:

- Changing annotations due to running a GeneSpider
- Adding genes due to loading an experiment
- Adding homology tables
- Adding notes to a gene
- Changing the map location associated with a gene due to the Problems with the Map Location Annotation window (from running the GeneSpider or from the regulatory sequence search window)
- Creating the genome
- Adding or removing genes in the Master Table of Genes window
- Adding, removing, or changing annotation entries in the Master Table of Genes window
- Adding or removing annotation fields in the Master Table of Genes
- Adding or updating the sequence in the Genome Inspector
- Publishing a genome to Signet (for Signet genomes only)
- Coping a genome from Signet (local genomes only)
- Updating a genome due to an external program (local genomes only)

This tab contains the following elements:

- **Operation**—Displays the list of actions associated with the genome.These operations are shown in Table 6-3.
- **Date**—Displays the date on which the operation occurred.
- **Results**—Describes the results of the operation.

Table 6-3 Genome Inspector: History Tab

Operation Label	Results
Run GeneSpider using GenBank-or-Locus Link-or-UniGene (displays the name of the database you spidered)	• Number of annotations changed • Number of genes with annotation changes
Run GeneSpider using Silicon Genetics (list of databases used, in order of priority)	• Number of annotations changed • Number of genes with annotation changes
Load Samples: Name-of-Sample 1, Name-of-Sample 2, …	• Number of genes added
Build Homology Tables	• Number of homology tables added • Number of homology tables replaced

Table 6-3 Genome Inspector: History Tab (Continued)

Add Notes to systematic-name-of-gene (common-name-of-gene)	• One annotation changed
Create Genome	• Number of genes • Number of annotation fields • Full nucleotide sequence or blank
Edit Master Table of Genes	• Number of genes added • Number of genes removed • Number of annotations changed • Number of annotation fields added • Number of annotation fields removed • Number of genes with annotation changes
Update genome with Name-of-External-Program (This entry gets recorded for local genomes only)	• Number of genes added • Number of annotations changed • Number of genes with annotation changes • Number of homology tables added • Number of homology tables replaced • Full nucleotide sequence added, full nucleotide sequence updated, or blank
Update genome from Name-of-Signet (This operation occurs when the user copies the genome from Signet; this entry only gets recorded for local genomes.)	• Number of genes added • Number of annotations changed • Number of genes with annotation changes • Number of homology tables added • Number of homology tables replaced • Full nucleotide sequence added, full nucleotide sequence updated, or blank
Update genome by user-name (This operation occurs when a user updates the genome on Signet from their local genome; this entry only gets recorded for Signet genomes.)	• Number of genes added • Number of annotations changed • Number of genes with annotation changes • Number of homology tables added • Number of homology tables replaced • Full nucleotide sequence added, full nucleotide sequence updated, or blank
Update Sequence	• Full nucleotide sequence updated
Add Sequence	• Full nucleotide sequence added
Extant Genome	• Number of genes • Number of annotation fields • Number of homology tables • Full nucleotide sequence or blank

Table 6-3 Genome Inspector: History Tab (Continued)

Remove homology table name-of-homology-table	• Remove homology table from genome 1 to genome 2
Download genome from name-of-Signet	• Number of genes • Number of annotations • Number of homology tables • Full nucleotide sequence or blank
Upload genome by user-name	• Number of genes • Number of annotations • Number of homology tables • Full nucleotide sequence or blank
Save UniGene Cluster IDs	• Number of annotations changed • Number of genes with annotation changes

Extant Genome

When upgrading from a version of GeneSpring prior to 7, all the genomes that were present at the time of the upgrade, will list a special operation in the History tab, called "Extant genome" and a date. This operation indicates that conversion of the genome from a previous version to the new format in version 7, and the date represents the conversion date, not the original date of creation of the Genome.

GenBank/EMBL Files Tab

The GenBank/EMBL Files tab (Figure 6-6) appears in the Genome Inspector only if GenBank or EMBL files were originally used to define the genome.

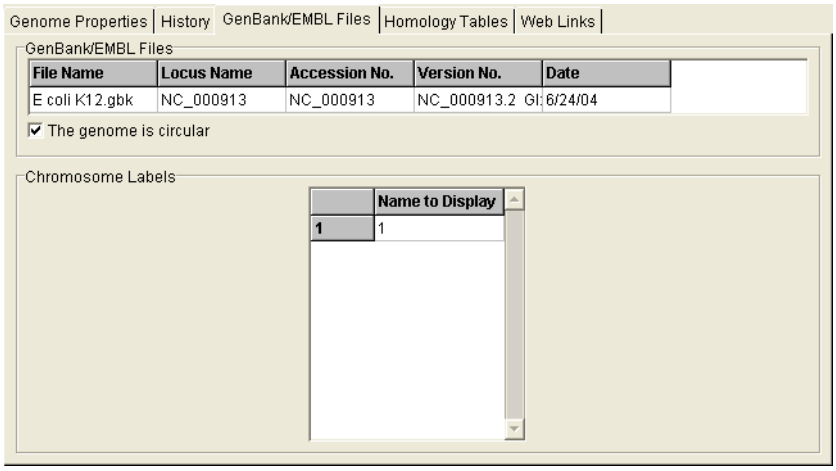


Figure 6-6 Genome Inspector Window: GenBank/EMBL Files Tab

GenBank/EMBL Files Table

The GenBank/EMBL Files table contains the following elements:

- **File Name**—Displays the file name for all GenBank or EMBL files used to define the genome.
- **Locus Name**—Displays the locus name for the GenBank entry in the file.
- **Accession No.**—Displays the accession number of the GenBank entry in the file.
- **Version No.**—Displays the version number for the GenBank entry and the GI Identifier in the file.
- **Date**—Displays the date the GenBank entry was last modified (not the date the file was last used in GeneSpring).

Chromosomes Table

The Chromosomes table lets you provide a name for your chromosomes. It lists each chromosome defined in the GenBank files or in the Map Location column of the Master Table of Genes. Information you enter in this table affects both the display of the Physical Position View and the list of chromosome names that appear in the various Search windows.

This table contains the following elements:

- **Name to Display**—Lets you enter the name for the chromosome.

  Note: This field is used with, or without, chromosome maps.

- **Circular**—Lets you designate whether the chromosome is circular.If your organism is a circular genome (such as a bacterium, plasmid, or virus), select **Yes** in the second box. This tells GeneSpring to display your genome as a circle in the physical position display. If your organism does not have a circular genome, leave **No**. selected.

  This option is only available if the genome has one chromosome.

Sequence Tab

The Sequence tab (Figure 6-7) lets you define the genomic sequence. This tab only appears if a genome has not been previously defined by GenBank or EMBL files, or if the sequence was defined during genome creation with seq files.

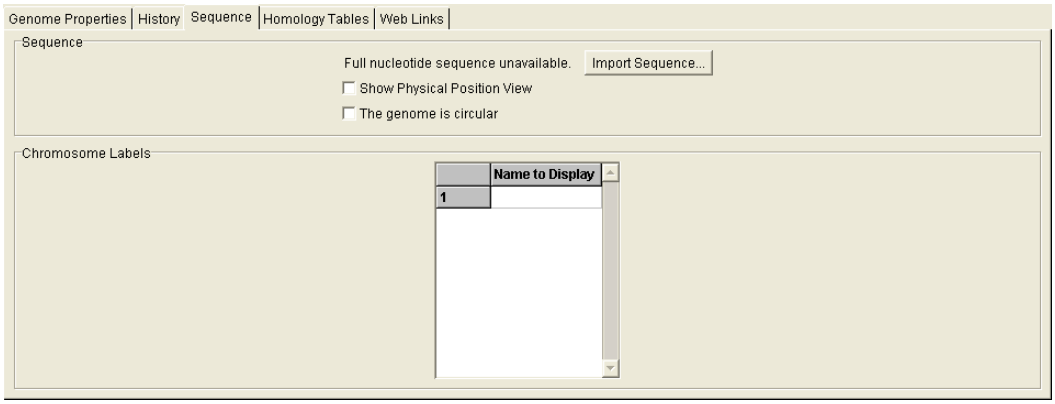


Figure 6-7 Genome Inspector Window: Sequence Tab

Sequence Box

- **Full nucleotide sequence unavailable**—Lets you import the full nucleotide sequence by clicking the Import Sequence button.
- **Full nucleotide sequence available. Last Updated: Month/Day/Year—**If a full nucleotide is available, this message appears and informs you of the last date the nucleotide sequence was updated.

  Note: These two options are mutually exclusive. If a sequence is already loaded, you will see the first option, **Full nucleotide sequence unavailable**, and you can load the sequence. If a sequence is already loaded, you will see the second option, **Full nucleotide sequence available**, and will be able to re-load sequence data.
- **Import Sequence**—Lets you import the nucleotide sequence.
- **Show Physical Position view**—Controls whether the Physical Position option appears in the View menu. This option is selected if the genome has chromosome maps and map locations that use the simple identifier format (example: “12q3.2”), or a full nucleotide sequence and map locations that use sequence location nomenclature, such as `<chromosome number>:<start_basepair> ..<end_basepair>` (example: “1:45378..574862”).
- **The genome is circular**—This check box is cleared if there is more than one chromosome.

Chromosome Labels Box

The Chromosome Labels box lets you name your chromosome. It lists all chromosomes mentioned in the GenBank files or in the map location column in the Master Table of Genes.

- **Name to Display**—Allows you to name the chromosome (even if the genome has chromosome maps).
- **Row Headers**—Displays the default names or numbers for the chromosomes.

Web Links Tab

The Genome Inspector: Web Links tab (Figure 6-8) lets you add web links to a genome.

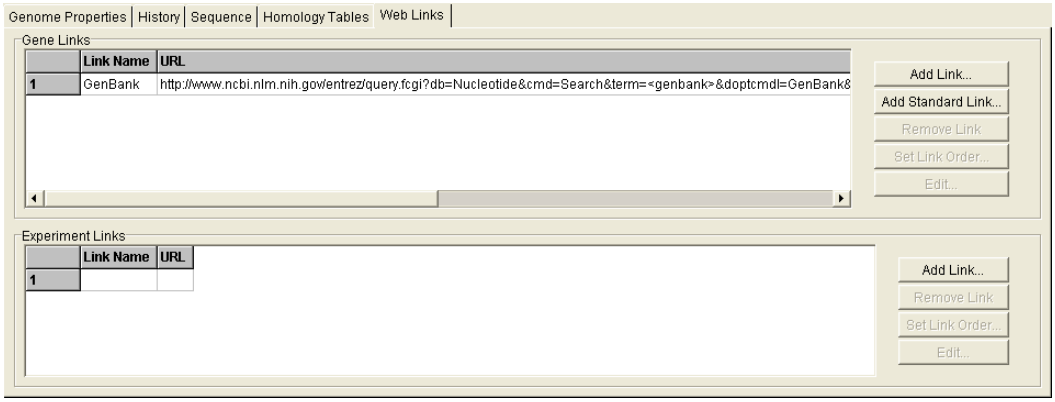


Figure 6-8 Genome Inspector Window: Web Links Tab

Gene Links Table

The Gene Links box lets you define the web links that are available in the Gene Inspector window for this genome.

- **Link Name**—Displays the name GeneSpring will use to refer to a particular web link.
- **URL**—Displays the URL GeneSpring will open when users select this link.
- **Add Link Button**—Displays the Add Link window where you can type in a gene link.
- **Add Standard Link Button**—Displays the Add Standard Link window where you can select gene links from a list.
- **Remove Link Button**—Deletes the selected web link.
- **Set Link Order Button**—Displays the Set Link Order window. This button is enabled whenever web links are present in the table.
- **Edit Button**—Displays the Edit Link window. This option is only enabled when a row is selected.

Experiment Links Table

The Experiment Links table lets you define the web links that are available in the Condition Inspector window for this genome. It contains the following elements:

- **Link Name**—Displays the name of the web link.
- **URL**—Displays the URL for the web link.
- **Add Link Button**—Displays the Add Link window where you can type in a web link.
- **Remove Link Button**—Deletes the selected web link.
- **Set Link Order Button**—Displays the Set Link Order window.
- **Edit Button**—Displays the Edit Link window. This option is only enabled when a row is selected.

Homology Tables Tab

The Homology Tables tab (Figure 6-8) displays all the homology tables that are available for this genome, the date on which they were created or updated, the number of homolog found for the genes in this genome, and the percentage of this genome that have homolog.

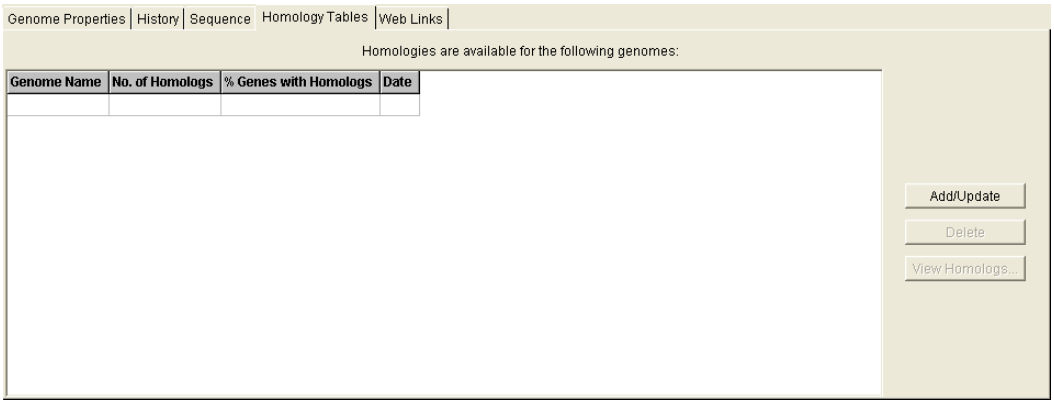


Figure 6-9 Genome Inspector Window: Homology Tables Tab

- **Genome Name**—Displays the name of the genome for which the homology table contains the homolog information.
- **No. of Homologs**—Displays the total number of genes in the inspected genome that have one or more homologs in the listed genome.
- **% Genes with Homologs**—Displays the percent of genome the user is inspecting that has homologs in the listed genome.
- **Date**—Displays the date on which the homology table was last updated.
- **Add/Update Button**—Displays the Build Homology Tables window.
- **Delete Button**—Deletes the selected homology table in the Homology Tables window.
- **View Homologs Button**—Displays the selected homology table.

Opening the Genome Inspector

To open the Genome Inspector:

1. Do one of the following:
   - Select **File > Genome Manager**, select the genome you want to inspect, and then click the **Inspect** button.
   - Select **Annotations > Edit Genes and Annotations**, select a genome you want to inspect, and then click the **Inspect Genome** button.

     For more information on the Edit Genes and Annotations window, go to “Managing Genes and Annotations” on page 6-35.

Editing Gene Properties

To edit gene properties

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.
2. Click the **Genome Properties** tab.
3. In the **Author** field, enter your name.
4. In the **Research** Group field, enter the name of your research group.
5. In the **Notes** box, enter your notes.
6. To edit **Gene Labels**, do the following:
   a. In the **Primary Gene Label** list, select the label you want.
   b. In the **Secondary Gene Label** list, select the label you want.
   c. To reset the labels, click the **Default Gene Labels** button.
   d. Click the **OK** button.

Viewing Genome History

The History table tracks all the changes that have been made to the genome. This includes:

- Running the GeneSpider
- Adding or deleting genes
- Building homology tables
- Modifying annotations
- Updating nucleotide sequences
- Updating data from Signet

To view genome history:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.
2. Click the **History** tab.
3. Go to "History Tab" on page 6-13 for a description of the parameters.

Importing Genomic Sequences

This procedure explains how to import genomic sequences into GeneSpring using the Genome Inspector. The file format that you import must be in seq format. Go to "Sequence File Format" on page 4-12 for more information about this file format.

To import a genomic sequence:

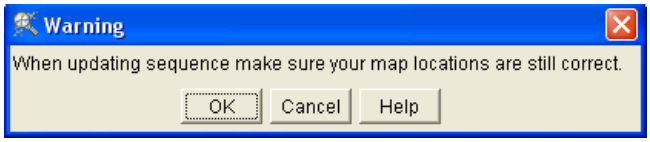
1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Sequence** tab.

3. Click the **Import Sequence** button.

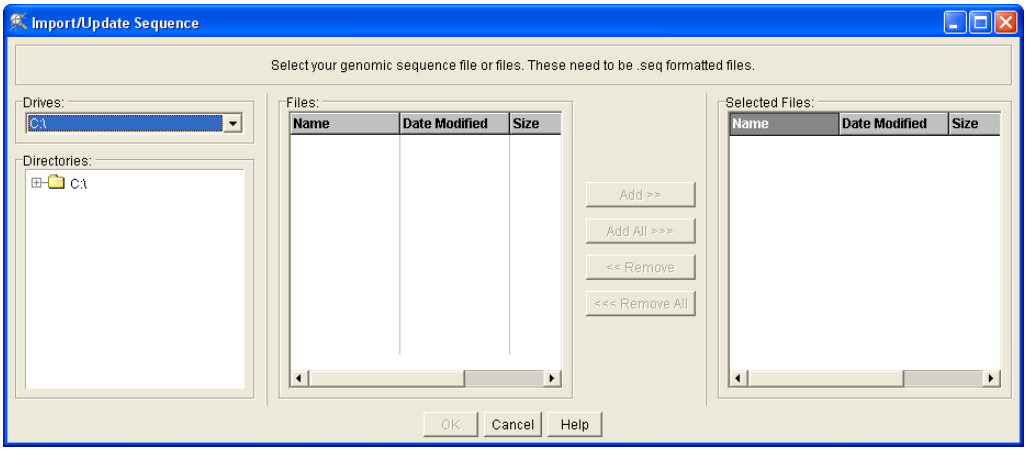
   A warning message opens.



   When you load the genomic sequence information, you need to make sure that the map locations that you define are correctly mapping to the actual genomic sequence you are loading.

4. Click the **OK** button to confirm that your map locations are correct.

   The Import/Update Sequence window opens.



5. In the **Drive** menu, select the drive you want.

6. In the **Directories Navigator**, select the directory you want.

7. In the **Files Navigator**, do the following:

   - To sort by columns, click the column title.
   - To select one file, single-click the file.

- To select several files at once, press Shift+click or Ctrl+click.
- To deselect a file, press Ctrl+click.

8. Click the **Add>>** button to add the selected files, or click the **Add All>>** button to all the files.

   The selected files appear in the Sequence Files box.

9. Click the **OK** button.

Building Homology Tables

Homology tables allow you to compare the expression results from one genome or array to the results from a different genome or array. Once an homology table is created, you can "translate" the gene list or experiment from one genome to a gene list of the corresponding genes in the other genome. This is useful if you have performed similar experiments on different chips or technologies and want to compare the behavior of a set of genes on the two technologies.

The Homology tool automates the process of building homology tables for certain organisms. Currently, a limited list of organisms is available. Using this tool, homologies can be made between any pair of organisms that are included in both HomoloGene and UniGene. Within-genome homologies are based solely on UniGene Cluster ID or LocusLink Locus ID.

This section provides a brief overview of how to build a new homology table. Go to "Working with Homology Tables" on page 12-23 for more information.

To build a homology table:

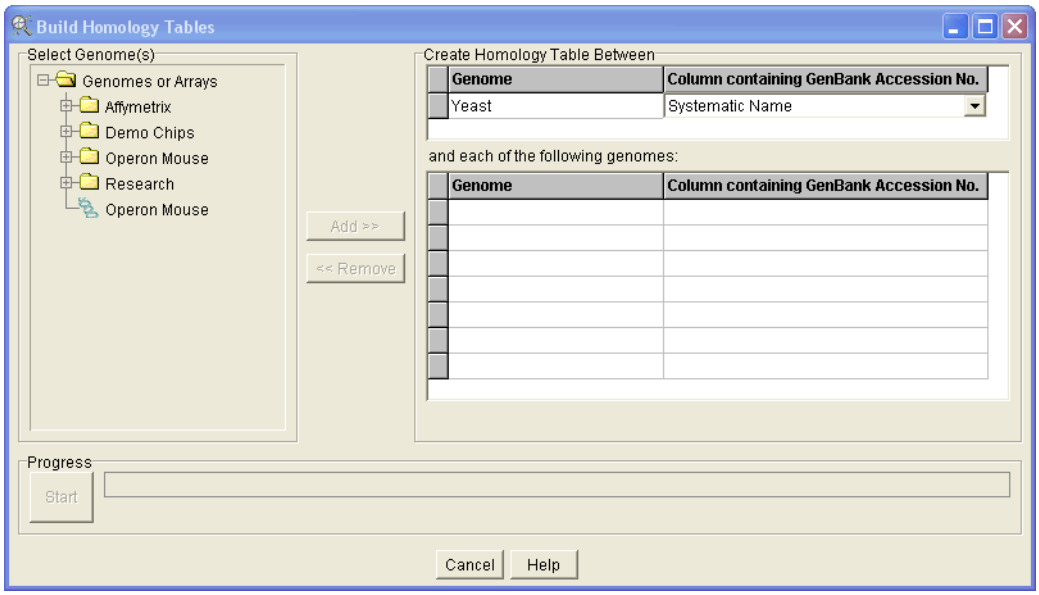
1. Do one of the following:

   - Open the **Genome Inspector**, click the **Homology Tables** tab, and then click the **Add/Update** button.

     Go to "Opening the Genome Inspector" on page 6-20 for instructions on opening the Genome Inspector.

   - Select **Annotations > Build Homology Tables**.

   The Build Homology Tables window opens.



2. From the menu in the **Column Containing GenBank Accession No.**, select the appropriate column that contains the GenBank (or EMBL) accession number for the selected genome.
3. In the **Navigator**, select the genome you want and click **Add**.

   The genome is added to the lower table on the right side of the window.
4. From the menu next to the newly added genome, select the name of the column containing the genome's GenBank Accession Number.
5. Repeat steps 2 through 4 for each genome you want to add.
6. Click the **Start** button.

   This process can take several hours to complete. If the initially selected genome does not have GenBank Accession Numbers, an error message appears. If a selected genome is not on Homologene, you will receive an error message after the Homology tool has finished running.
7. When prompted, specify whether or not to save the UniGene Cluster IDs.

   When you choose to select the UniGene ID's, they will be saved in a annotation column called "Unigene". Any data that was previously in the Unigene column will be overwritten.

   The resulting homology tables are saved in the originally selected genome.

Managing Gene Links

This section explains how to manage gene links in the Genome Inspector: Web Links Tab using the Add Link, Add Standard Link, Remove Link, Set Link Order, and Edit buttons. It covers the following topics:

- About Gene Links
- Adding Custom Gene Links
- Adding Standard Gene Links
- Setting the Gene Link Order
- Removing Gene Links
- Editing Gene Links

About Gene Links

GeneSpring now has the ability to use different terms from the same annotation entry, provided that these terms are separated in some way (comma separated lists and semicolon separated lists are the most common). All current links will continue to work as they did in previous releases.

To enable this feature, GeneSpring has extended the hypertext link format as follows:

- To cycle though a semicolon separated list of common names use:

  `<Common:([^;]*)>`

- To cycle though a comma separated list of Synonyms use:

  `<Synonym:([^,]*)>`

where `Synonym` indicates the name of the annotation and the comma (`,`) indicates the separation character.

Adding Custom Gene Links

To add a custom gene link:

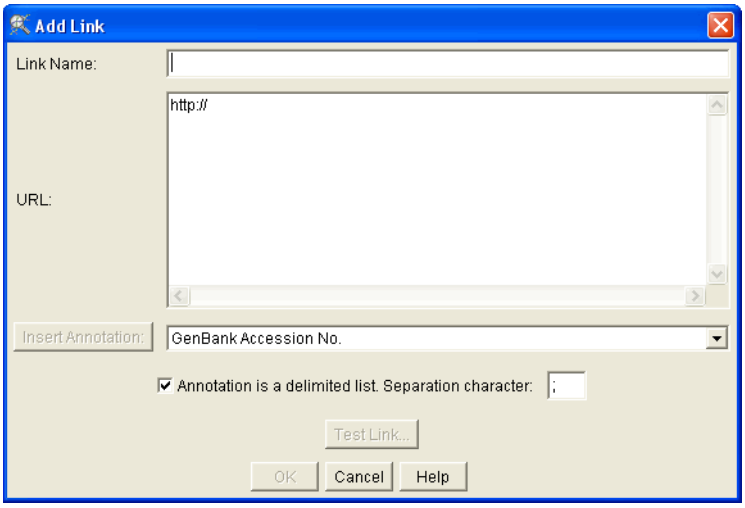
1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Web Links** tab.

3. Click the **Add Link** button in the Gene Links table.

   The Add Link window opens.



4. In the **Link Name** field, enter the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Gene Inspector.
5. In the **URL** field, paste the URL you want for a particular web database.

   For example:

   ```
   http://www.google.com/search?sourceid=
   navclient&q=<COMMON>:([^:]*)>+human+OR+sapiens
   ```

   will use the common name to query Google.
6. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Annotation** menu, select the annotation you want to add.
   c. To ensure that the annotation includes the regular expression for "loop over a delimited list" (where the delimiter is the character given in the Separation Character box), select the **Annotation is a delimited list** check box.

      The Separation Character box defaults to a semicolon.
7. Click the **Insert Annotation** button.

   The annotation is placed in the URL field with the correct punctuation. The default selection is the GenBank Accession Number or the Systematic Name (if the GenBank Accession Number is unknown).
8. To test the link, click the **Test Link** button.

   If the test fails, GeneSpring will be unable to open the link. Be sure you have entered a valid link.
9. Click the **OK** button.

10. Do one of the following:

- To add more web links, return to “Managing Gene Links” on page 6-25.
- If you are done adding web links, click the **OK** button.

Adding Standard Gene Links

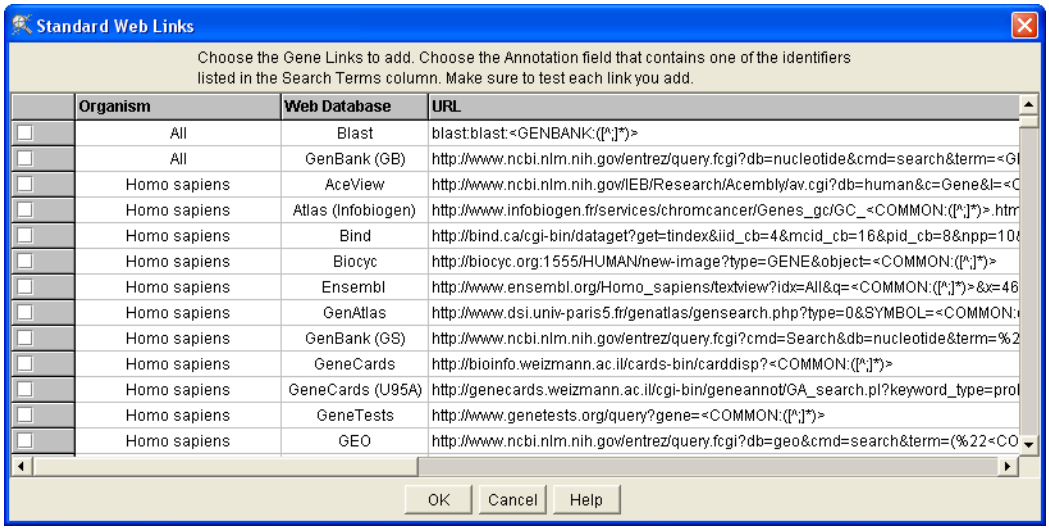
To add a standard gene link:

1. Open the **Genome Inspector**.

   Go to “Opening the Genome Inspector” on page 6-20 for instructions.

2. Click the **Web Links** tab.
3. Click the **Add Standard Link** button in the Gene Links table.

   The Add Standard Link window opens.


Standard Web Links
Choose the Gene Links to add. Choose the Annotation field that contains one of the identifiers listed in the Search Terms column. Make sure to test each link you add.
Organism
Web Database
URL
All Blast blast:blast:<GENBANK:([^;]*)>
All GenBank (GB) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&cmd=search&term=<G
Homo sapiens AceView http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&c=Gene&l=<C
Homo sapiens Atlas (Infobiogen) http://www.infobiogen.fr/services/chromcancer/Genes_gc/GC_<COMMON:([^;]*)>.htm
Homo sapiens Bind http://bind.ca/cgi-bin/dataget?get=tindex&iid_cb=4&mcid_cb=16&pid_cb=8&npp=10
Homo sapiens Biocyc http://biocyc.org:1555/HUMAN/new-image?type=GENE&object=<COMMON:([^;]*)>
Homo sapiens Ensembl http://www.ensembl.org/Homo_sapiens/textview?idx=All&q=<COMMON:([^;]*)>&x=46
Homo sapiens GenAtlas http://www.dsi.univ-paris5.fr/genatlas/gensearch.php?type=0&SYMBOL=<COMMON:
Homo sapiens GenBank (GS) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=nucleotide&term=%2
Homo sapiens GeneCards http://bioinfo.weizmann.ac.il/cards-bin/carddisp?<COMMON:([^;]*)>
Homo sapiens GeneCards (U95A) http://genecards.weizmann.ac.il/cgi-bin/geneannot/GA_search.pl?keyword_type=prot
Homo sapiens GeneTests http://www.genetests.org/query?gene=<COMMON:([^;]*)>
Homo sapiens GEO http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo&cmd=search&term=(%22<CO
OK
Cancel
Help

This window lets you easily import pre-made web links. It contains the following elements:

- **Organism**—Displays the Latin names for the organism, or the term, “Various,” if the link can be applied to multiple organisms. For other Latin names, go to:
  - http://genome-www5.stanford.edu/cgi-bin/SMD/listMicroArrayData.pl?tableName=organism
  - http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi?view=platforms
- **Web Database**—Displays the names of the links as they will appear on buttons in the Gene Inspector window.
- **Column Headers**—Clicking on a column header sorts the table based on the entries.

- **URL**—Displays the URL for the web link.

  For a list of standard gene links that you can add, go to "Standard Gene URLs" on page 4-16.

- **Search Terms**—Provides examples of search terms that can be used in the Annotation list.

4. In the **Organism** column, locate the gene links you want to add.
5. Select the check box next to the organism(s) you want.

   Each check box you select will add a web link to the gene links table in the Genome Inspector: Web Links Tab.
6. In the **Annotation** list, select the annotation you want to use. The annotation should contain one of the types of annotations listed in the Search Terms column.

   Note: If you don't see the Annotation list, use the scroll bar to scroll to the right side of the window.
7. Click the **OK** button.
8. Do one of the following:

   - To add more web links, return to "Managing Gene Links" on page 6-25.
   - If you are done adding web links, click the **OK** button.

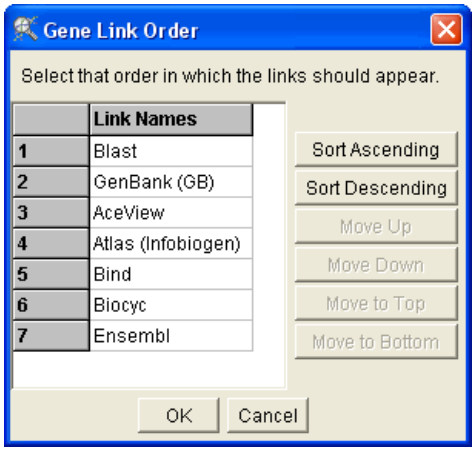
Setting the Gene Link Order

To set the gene link order:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.
2. Click the **Web Links** tab.
3. Click the **Set Link Order** button in the Gene Links table.

   The Gene Link Order window opens.



4. Do the following:
   - To sort the links in ascending order, click the **Sort Ascending** button.
   - To sort the links in descending order, click the **Sort Descending** button.
   - To move the selected link up one row, click the **Move Up** button.
   - To move the selected link down one row, click the **Move Down** button.
   - To move the selected link to the top of the column, click the **Move to Top** button.
   - To move the selected link to the bottom of the column, click the **Move to Bottom** button.
5. Click the **OK** button.
6. Do one of the following:
   - To add more web links, return to "Managing Gene Links" on page 6-25.
   - If you are done adding web links, click the **OK** button.

Removing Gene Links

To remove a gene link:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.
2. Click the **Web Links** tab.
3. Select a link to remove in the Gene Links table.
4. Click the **Remove Link** button.
5. Do one of the following:
   - To add more web links, return to "Managing Gene Links" on page 6-25.
   - If you are done adding web links, click the **OK** button.

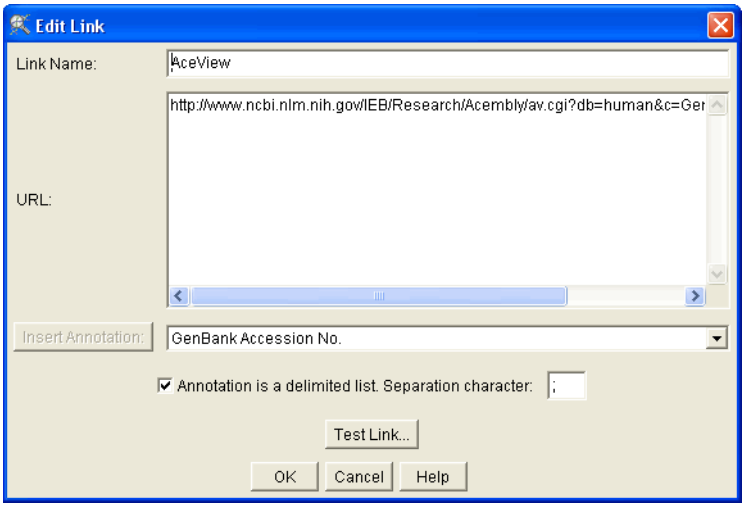
Editing Gene Links

To edit a gene link:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.
2. Click the **Web Links** tab.
3. Select a link to edit in the Gene Links table.
4. Click the **Edit** button.

   The Edit Link window opens.



5. In the **Link Name** field, edit the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Gene Inspector.
6. In the **URL** field, edit the URL you want for a particular web database.
7. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Annotation** menu, select the annotation you want to add.
   c. To ensure that the annotation includes the regular expression for "loop over a delimited list" (where the delimiter is the character given in the Separation Character box), select the **Annotation is a delimited list** check box.

      The Separation Character box defaults to a semicolon.
8. Click the **Insert Annotation** button.

   The annotation is placed in the URL field with the correct punctuation. The default selection is the GenBank Accession Number or the Systematic Name (if the GenBank Accession Number is unknown).
9. To test the link, click the **Test Link** button.

   If the test fails, GeneSpring will be unable to open the link. Ensure that the link you enter is valid.
10. Click the **OK** button.
11. Do one of the following:
    - To add more web links, return to "Managing Gene Links" on page 6-25.
    - If you are done adding web links, click the **OK** button.

Managing Experiment Links

This section explains how to manage experiment links in the Genome Inspector: Web Links Tab using the Add Link, Remove Link, Set Link Order, and Edit buttons. It covers the following topics:

- About Experiment Links
- Adding Experiment Links
- Setting the Experiment Link Order
- Removing Experiment Links
- Editing Experiment Links

About Experiment Links

GeneSpring now has the ability to use different terms from the same annotation entry, provided that these terms are separated in some way (comma separated lists and semicolon separated lists are the most common). All current links will continue to work as they did in previous releases.

To enable this feature, GeneSpring has extended the hypertext link format as follows:

- To cycle though a semicolon separated list of common names use:

  `<Common:([^;]*)>`

- To cycle though a comma separated list of Synonyms use:

  `<Synonym:([^,]*)>`

where `Synonym` indicates the name of the annotation and the comma (`,`) indicates the separation character.

Adding Experiment Links

To add an experiment link:

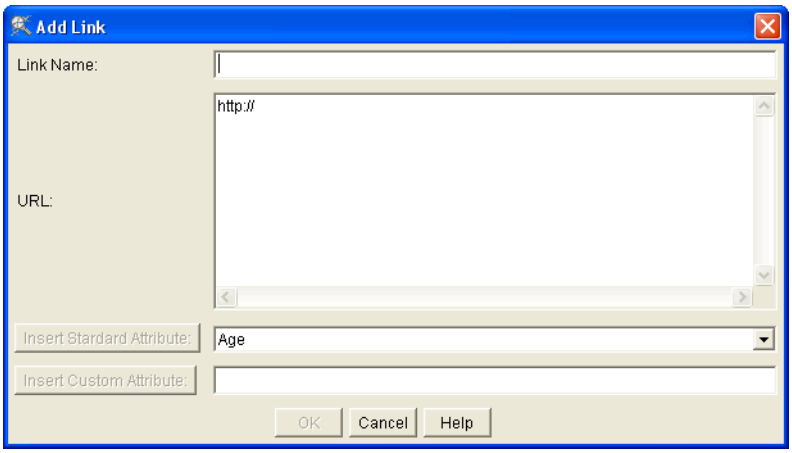
1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Web Links** tab.

3. Click the **Add Link** button in the Experiment Links table.

   The Add Link window opens.



4. In the **Link Name** field, enter the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Condition Inspector.
5. In the **URL** field, paste the URL you want for a particular web database.

   For example:

   `http://www.ncbi.nih.org.gov/geo/query/acc.cgi?acc=<GEO IDENTIFIER>`

   will query the GEO database using the sample attribute or parameter of the GEO IDENTIFIER.
6. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Standard Attribute** menu, select the annotation you want to add.
   c. Click the **Insert Standard Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
7. To add a custom annotation to the URL field, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Custom Attribute** menu, enter any kind of annotation you want.
   c. Click the **Insert Custom Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
8. Click the **OK** button.
9. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 6-31.
   - If you are done adding web links, click the **OK** button.

Setting the Experiment Link Order

To set the experiment link order:

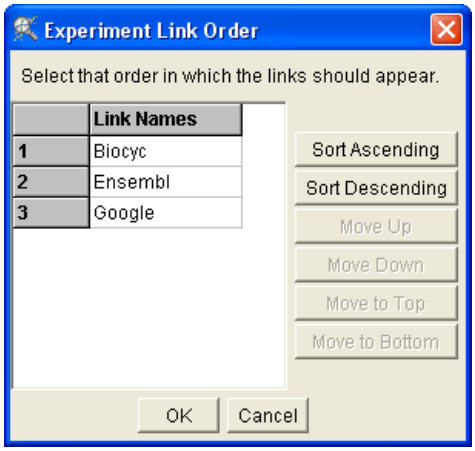
1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Web Links** tab.

3. Click the **Set Link Order** button in the Experiment Links table.

   The Experiment Link Order window opens.



4. Do the following:

   - To sort the links in ascending order, click the **Sort Ascending** button.
   - To sort the links in descending order, click the **Sort Descending** button.
   - To move the selected link up one row, click the **Move Up** button.
   - To move the selected link down one row, click the **Move Down** button.
   - To move the selected link to the top of the column, click the **Move to Top** button.
   - To move the selected link to the bottom of the column, click the **Move to Bottom** button.

5. Click the **OK** button.

6. Do one of the following:

   - To add more web links, return to "Managing Experiment Links" on page 6-31.
   - If you are done adding web links, click the **OK** button.

Removing Experiment Links

To remove an experiment link:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Web Links** tab.
3. Select a link to remove in the Experiment Links table.
4. Click the **Remove Link** button.
5. Do one of the following:
   - To add more web links, return to "Managing Experiment Links" on page 6-31.
   - If you are done adding web links, click the **OK** button.

Editing Experiment Links

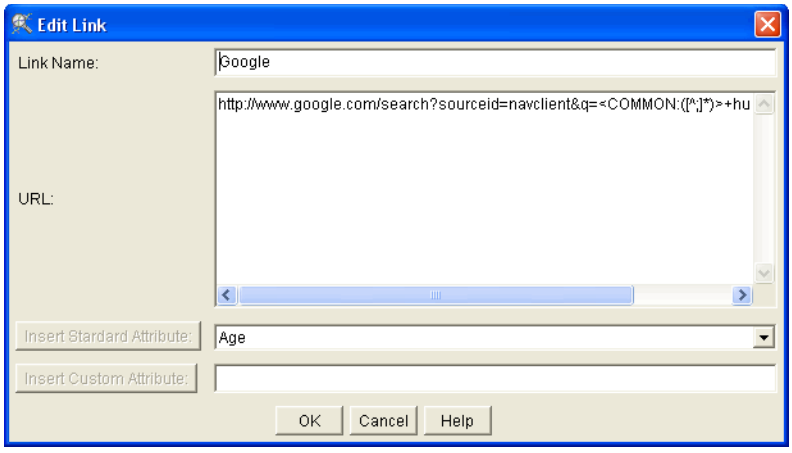
To add an experiment link:

1. Open the **Genome Inspector**.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions.

2. Click the **Web Links** tab.
3. Select a link to edit in the Experiment Link table.
4. Click the **Edit** button.

   The Edit Link window opens.



5. In the **Link Name** field, edit the name of the link GeneSpring will use to refer to this particular link. This name will appear in the Condition Inspector.
6. In the **URL** field, edit the URL you want for a particular web database.

7. To replace the search terms in the URL with the name of the general annotation fields, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Standard Attribute** menu, select the annotation you want to add.
   c. Click the **Insert Standard Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
8. To add a custom annotation to the URL field, do the following:
   a. Place the cursor in the URL field where you want to add the annotation.
   b. In the **Insert Custom Attribute** menu, enter any kind of annotation you want.
   c. Click the **Insert Custom Attribute** button.

      The annotation is placed in the URL field with the correct punctuation.
9. Click the **OK** button.
10. Do one of the following:
    - To add more web links, return to "Managing Experiment Links" on page 6-31.
    - If you are done adding web links, click the **OK** button.

Managing Genes and Annotations

In prior versions of GeneSpring versions, you had to manually modify the genome files (*_annotations.txt) in order to add custom annotations to your genomes. In GeneSpring 7, you can use the Master Table of Genes Editor to edit genes and create custom annotations.

This section explains how to use the Master Table of Genes Editor and covers the following topics:

- Master Table of Genes
- Master Table of Genes Editor Window
- Which Annotations are Retrieved?
- Opening the Master Table of Genes Editor
- Editing Columns
- Creating Columns for EBI Submissions
- Sorting Data
- Adding Genes
- Deleting Genes
- Importing Genes and Annotations
- Editing Annotations
- Finding Genes
- Finding and Replacing Text

- Copying Data to the Clipboard
- Saving Data to a File
- Adding Other Genomic Elements

Master Table of Genes

In GeneSpring, a genome includes all the genes on your chip or array. When you create a genome from experiment data, GeneSpring creates a genome on-the-fly, based on genes in your experiment data files. This means that, unlike a genome created with the Import Genome Wizard, this genome has no annotations and no means of obtaining annotations from public databases.

Master Table of Genes Editor Window

You use the Master Table of Genes Editor (Figure 6-10) to add, edit, or remove genes and annotations from the Master Table of Genes.

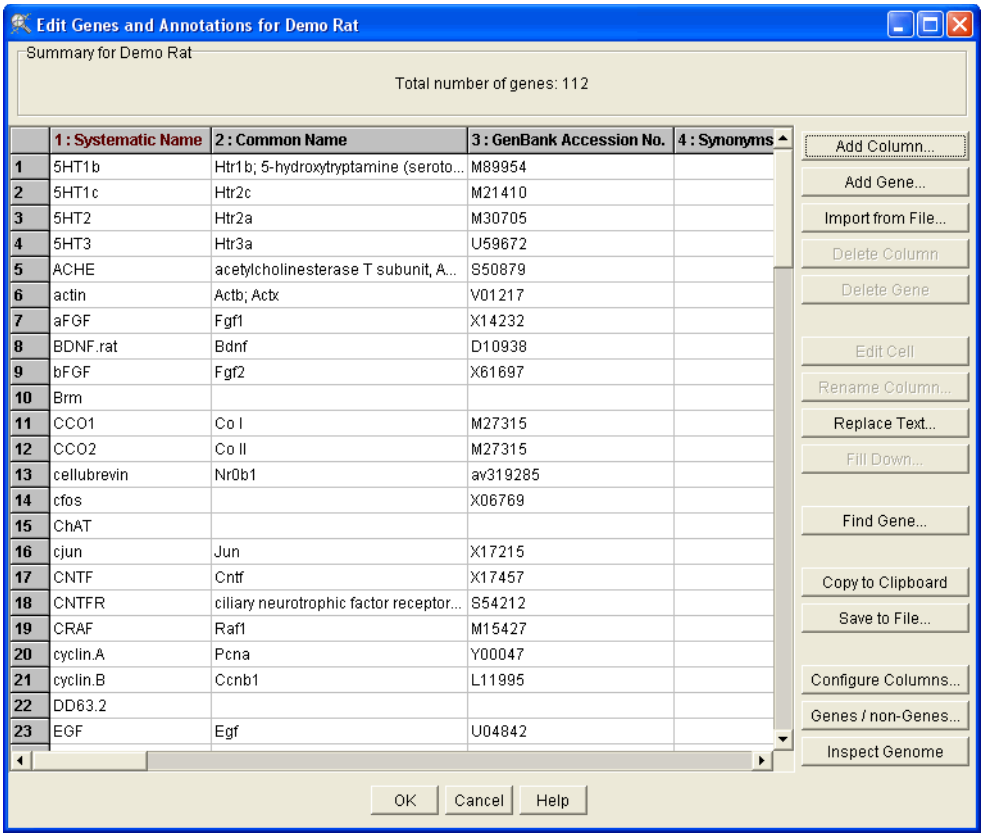


Figure 6-10 Master Table of Genes Editor Window

This section describes the elements in the Master Table of Genes Editor.

Summary Box

The Summary box at the top of the window displays the following information:

- **Title**—Display the genome name.
- **Total number of genomic elements**—Displays the total number of genes and other genomic elements in the genome.
- **Total number of genes**—Displays the number of rows whose Genomic Element Type is “gene”.
- **Total number of other genomic elements**—Displays the number of rows whose Genomic Element Type is “Not a Gene”. Items that fall into this category include genomic telomeres, CpG islands, and other regions on the genome that usually are not transcribed. This entry is excluded if the number of rows with “Not a Gene” is zero.

Table

The table can show any of the annotations in the Master Table of Genes files. It can also include a number of columns that are not currently part of the Master Table of Genes files, such as other genomic elements, spider errors, data added, and date modified.

Columns

The column titles include the number of the annotation column, followed by a colon. If columns are being hidden, the numbers will jump because the annotation numbers are being hidden.

The number associated with the annotation is determined from it’s placement in the Configure Columns window. Numbering occurs from left to right, starting with 1. Annotations that are longer then the cell are truncated with ellipses. Clicking on a column heading sorts the rows in ascending or descending order.

The Master Table of Genes contains a number of columns that have special meanings in GeneSpring. These include annotation columns with names that have specific values, as well as a number of generic annotation columns in which the column names do not have special values.

Identifier Columns

There are four columns that are used as the so-called identifier columns: “Systematic Name”, “Common Name”, “GenBank”, and “Synonym”.

- **Systematic Name**—The unique identifier for the gene in this genome or array. It is recommend that the gene’s systematic name be used to label the gene’s expression values in your experiment data files.
- **Common Name**—An alternative way of referring to this gene. Genes are not required to have a common name, and common names do not have to be unique, although duplicated common names may lead to confusion if the common name is how the gene is referred to in the experiment files. Usually, the common name annotation column can be used to store the HUGO gene symbol or some other official gene identifier.

- **GenBank—**The GenBank accession number for this gene, if known. If the GenBank identifiers for your genes were not used as either their Systematic Names or Common Names, then including the GenBank Accession Number in this field allows you to update the information about this particular gene directly from GenBank.
- **Synonym**—This column allows for other names to be entered for the genes. Multiple names should be separated by semicolons (;).

GeneSpider Columns

The next set of special columns are those that are created and edited by the GeneSpider If these columns do not yet exist, GeneSpider creates them. If these columns do exist, GeneSpider updates them. For information on which columns are retrieved by GeneSpider, go to "Which Annotations are Retrieved?" on page 6-41.

- **Map**—Sequence position of the gene on the chromosome. There are two formats for the Map column value. <`chromosome_number>:<start_bp>..<end_bp>` and `<position_id>.`

  The first format is usually used if the gene position in base pairs is known exactly, and with the completion of many genomes, this is true for many of the genes. An example can be "1:45378..58592". If the chromosome sequence is available, the genes will be mapped and the actual gene sequence can be retrieved for a gene.

  The second format is used when the exact position of the gene is not known, but an approximate position is known. An example of a cytogenic position map position is: "16q21.1".

  In this case there usually is a Chromosomal Map file, that defines the cytogenetic location and the coloring of the cytogenetic bands. For more information, see the section on Chromosome map files "Chromosomes Table" on page 6-17.
- **Description**—A description of this gene, if known.
- **EC number**—The Enzyme Commission (EC) number for this gene, if known.
- **Product**—The protein product coded for by this gene, if known. This information can be accessed when you use the Find Gene command.
- **Phenotype**—A description of the phenotype for this gene, if known.
- **Function**—A description of the function of this gene product, if known.
- **Keywords**—Keywords associated with this gene, if known. Separate keywords with semicolons. This information can be accessed when you use the Find Gene command.
- **PubMedID**—The Public Medline accession number, if known. Multiple identifiers should be separated by semicolons (;).
- **Type**—A result of the conversion from a gbk file to a master table of genes. It come from the GenBank column "feature type". For example, possible entries include: CDS, gene, terminator, rRNA.
- **Database reference (DBid)**—A specific field returned by the GeneSpider. There are dbxref entries in GenBank, and these entries provide database ID for other, non-GenBank databases, such as the SwissProt ID numbers. There may be multiple entries for each gene.

- **GO Biological Process**—The Gene Ontology Biological Process classification.
- **GO Molecular Function**—The Gene Ontology Molecular Function classification.
- **GO Cellular Component**—The Gene Ontology Cellular Component classification.
- **RefSeq**—The gene's National Center for Biotechnology Information (NCBI) Reference Sequence project identifier.
- **UniGene**—The gene's UniGene cluster identifier.

Date Columns

A next set of columns indicates the dates of addition and modification of the gene annotation.

- **Date Added—**The date a gene was added to GeneSpring.
- **Date Modified—**The date the gene was modified in GeneSpring.

Genomic Element Type Column

- **Genomic Element Type—**The cells in this column can only have two possible entries: Gene or Not a Gene.These are the entries that are designated as Other Genomic Elements in the Define Other Genomic Elements window. This means that the row will be ignored in Regulatory Sequence Searches and will be part of the All Genomic Elements gene list, but not part of the All Genes gene list. The cells in this column are not editable. They can only be modified from the Define Other Genomic Elements window.

Cells

You can edit most cells in the table by clicking the cell; however, the following columns are not editable:

- Systematic Name
- Date Modified
- Date Added
- Genomic Element Type (this cell can be set in the Define Other Genomic Elements window.)

Buttons

- **Add Column**—Adds an empty column to the table and displays the Add Annotation window (which lets you name the annotation).
- **Add Gene**—Displays the Add Gene window which lets you add a gene to the genome. The systematic name of the genome must be unique.
- **Import from File**—Displays the Import from File window enabling you to import genes and annotations from an outside file.

- **Delete Column**—Deletes the selected column or columns. The following columns, however, cannot be deleted:
  - Systematic Name
  - Common Name
  - GenBank Accession Number
  - Synonyms
  - Map
  - Date Modified
  - Date Added
  - Genomic Element Type
- **Delete Gene**—Deletes the selected row or rows.
- **Find Gene**—Displays the Find Gene window and searches the table for the user-entered gene.
- **Edit Cell**—Displays the Edit Annotation Value for the selected cell. This button is only enabled if one and only one cell is selected.
- **Rename Column**—Displays the Edit Annotation window that lets you rename an annotation. Annotations must have unique names.
- **Replace Text**—Displays the standard Replace window that lets you to search and replace text.
- **Fill Down**—Fills in all the selected cells with the text from the top selected cell. This button is only enabled when a single column is selected or when two or more cells in the column are selected. GeneSpring will fill all cells below the top cell with the values from the top cell. This tool enables you to quickly copy the contents from one cell to all cells below it and saves you time in entering information.
- **Sort**—Sorts the genes in ascending or descending order based on the entries in the selected column. Clicking on a column header sorts the column.
- **Copy to Clipboard**—Copies the entire table to the clipboard (irrespective of what is selected). This includes only the columns that are actually being displayed.
- **Configure Columns**—Displays the Configure Columns window that lets you modify what annotations are being displayed.
- **Genes/non-Genes**—Displays the Define Other Genomic Elements window that lets you define the Genomic Element Type (for example, create a list of Other Genomic Elements).
- **Inspect Genome**—Displays the Genome Inspector for this genome.
- **OK**—Saves the edits and closes the Master Table of Genes Editor.
- **Quit**—Closes the Master Table of Genes Editor without saving the edits.

Menu

Right-clicking on a cell in the Master Table of Genes Editor displays commands for performing different actions. Table 6-4 describes the commands that can appear.

Table 6-4 Master Table of Genes Editor: Menu

Command	Description
Cut	Cuts the text from the selected cell.
Copy	Copies the text from the selected cell.
Paste	Pastes the copied text to the selected cell.
Paste Transpose	Pastes the data on the clipboard and transposes the data in the process. When the data on the clipboard is orientated as a column, the data will be pasted in rows. Alternatively, when data on the clipboard is orientated in rows, the data will be pasted as a column.
Clear	Clears the text in the selected cell.
Delete Columns	Deletes the selected column. Note: This command cannot be used to delete the following annotations: • Systematic Name • Common Name • Synonym • GenBank Accession Number • Date Modified • Date Added • Genomic Element Type
Rename Column	Opens the Column Editor window.

Which Annotations are Retrieved?

Table 6-5 describes the annotation fields retrieved by GeneSpider from the various databases:

Table 6-5 Database Annotation Fields

Annotation	Silicon Genetics	GenBank	LocusLink	UniGene
Systematic name				
Common	X	X	X	X
Map	X	X	X	X
GenBank				
Synonym				
EC number	X	X	X	
Description	X	X		X
Product	X	X	X	

Table 6-5 Database Annotation Fields (Continued)

Phenotype	X	X	X	
Function	X	X		
Keywords	X	X		
PubMedID				
Type				
DBId	X	X		
GO biological process	X			
GO molecular function	X			
GO cellular component	X			
RefSeq	X			
UniGene	X			
Sequence	X	X		

Opening the Master Table of Genes Editor

To open the Master Table of Genes:

- Do one of the following:
  - Select **Annotations > Edit Genes and Annotations**.

    Note: You cannot open the Master Table of Genes Editor when a Homology Table window (for any genome) is open or when the GeneSpider is running.
  - Open the **Genome Inspector**, click the **Genome Properties** tab, and then click the **Edit Genes and Annotations** button.

    Go to "Opening the Genome Inspector" on page 6-20 for instructions.

Editing Columns

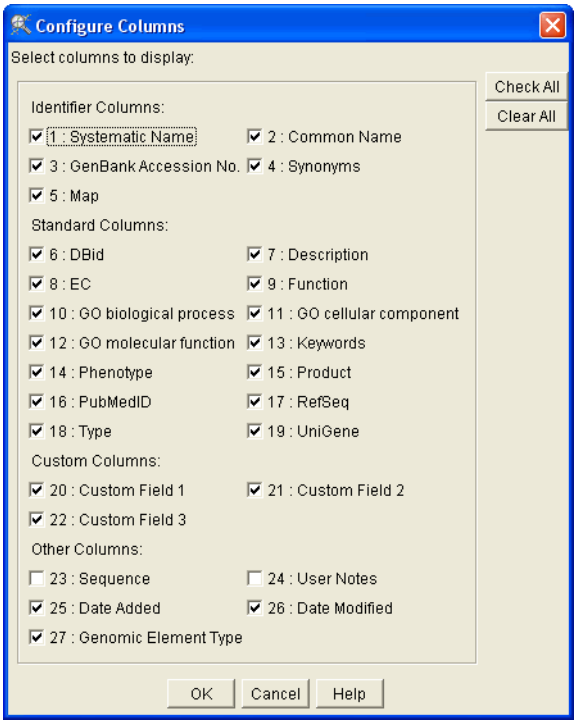
To edit columns:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.
2. To configure the columns, do the following:

   a. Click the **Configure Columns** button.

The Configure Columns window opens.



b. To enable all annotation columns to display, click the **Check All** button.

c. To remove all annotation columns from the view, click the **Clear All** button.

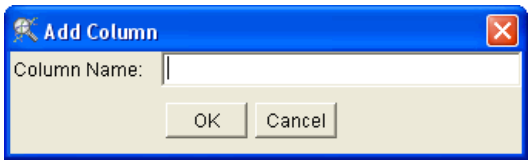
d. To select individual options, select the options you want.

e. Click the **OK** button.

3. To add an annotation column, do the following:

a. Click the **Add Column** button.

The Add Column window opens.



b. Enter a name for the column and click the **OK** button.

The annotation name must be unique to this genome. The name cannot exceed 200 characters.

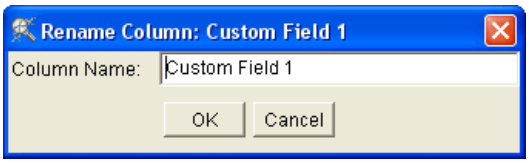
Warning: If you enter an annotation already used by GeneSpider (such as Common Name, Map, or EC) the next time you run GeneSpider, it will return annotations and fill in or overwrite entries in this column (depending on the GeneSpider settings).

4. To delete an annotation column, do the following:
   a. Select the column you want to delete and click the **Delete Column** button.

      A confirmation window opens.

   b. Click the **Yes** button to confirm.
5. To rename an annotation column, do the following:
   a. Select the column you want to rename and click the **Rename Column** button.

      The Rename Column window opens.



   b. Enter the new name for the column and click the **OK** button.

      The annotation name must be unique to this genome. The name cannot exceed 200 characters.

      Warning: If you enter an annotation already used by GeneSpider (such as Common Name, Map, or EC) the next time you run GeneSpider, it will return annotations and fill in or overwrite entries in this column (depending on the GeneSpider settings).
6. To fill down an annotation column with data from a cell, do the following:
   a. To repeat a value from the top cell to other cells below it, select the range of cells you want to fill.
   b. Click the **Fill Down** button.

      Note: This button is only enabled if multiple cells are selected within a single column. It fills in all the selected cells with the text from the top selected cell.

Creating Columns for EBI Submissions

If you plan to submit new data or updated information to the European Bioinformatics Institute (EBI)[1], and you want to include EBI CompositSequence names with your experiment MAGE-ML files, you must map the gene identifiers in the experiment MAGE-ML file with the identifiers used by EBI.

To do so, use the Master Table of Genes Editor to create a new column for the EBI CompositSequence names. For more information about exporting MAGE-ML files, go to "Exporting MAGE-ML Data" on page 17-10.

1. EBI is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL). The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences, and macromolecular structures.

Sorting Data

To sort data:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.

2. Click the column header on the column you want to sort.

   The data is sorted in ascending/descending order based on the entries in the selected column.On the first click, the data sorts in ascending order. On the second click, the data sorts in descending order, as follows:

   - Text entries are sorted in alphabetical order.
   - Numeric entries are sorted in numerical order.
   - Date entries (as for the Date Modified, and Date Added columns) are sorted in chronological order.

Adding Genes

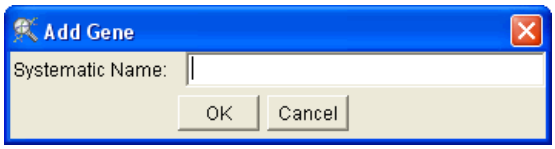
To add a gene:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.

2. Click the **Add Gene** button.

   The Add Gene window opens.



3. Enter a name for the gene and click the **OK** button.

   Note: Systematic Names are limited to 256 characters.

   - If the name is not unique, you will get an error message and a prompt to rename the gene.
   - If the name is a unique Systematic Name, but occurs in an existing gene's Common, Synonym, or GenBank Accession Number field, you will get a warning message.
   - If the name is unique, a new row is added to the bottom of the Edit Genes and Annotations table.

   GeneSpring automatically fills in the Systematic Name, Date Added, and Date Modified cells. The Genomic Element Type is set to Gene, and all other fields are empty.

Deleting Genes

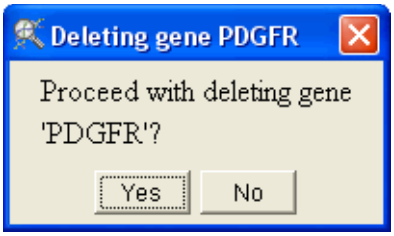
To delete a gene:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.

2. Click the **Delete Gene** button.

   A warning window opens.



3. Click the **Yes** button to confirm or **No** to cancel.

Importing Genes and Annotations

You can import both genes and annotations into an existing genome from a tab-delimited text file. This section contains the following topics:

- Starting the Import Genes and Annotations Process
- Fixing Duplicate Identifiers
- Fixing Long Gene Name Problems
- Completing the Import Genes and Annotations Process

Starting the Import Genes and Annotations Process

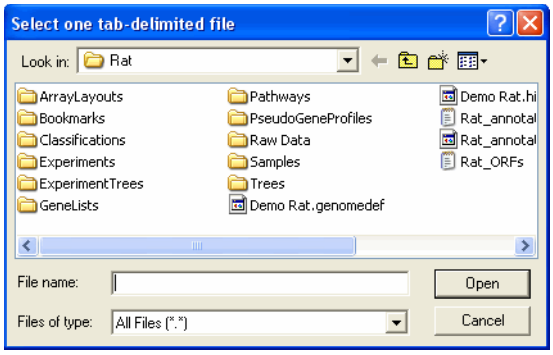
To import genes and annotations:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.

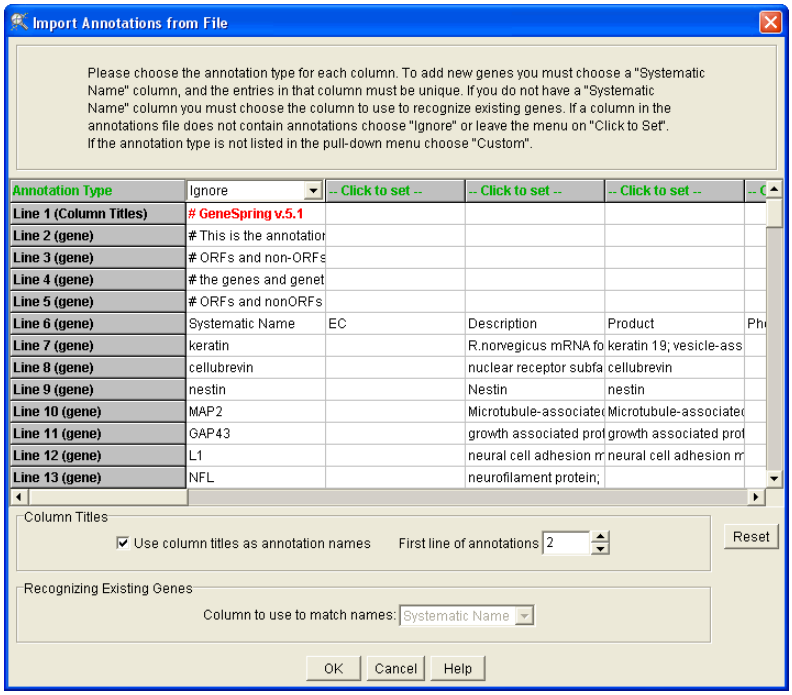
2. Click the **Import from File** button.

The Select One Tab-delimited File window opens.



3. In the **Look in** menu, click the drive, folder, or Internet location that contains the file you want to open.
4. In the **Folder** menu, locate and open the folder that contains the file.
5. Select the file, and then click the **Open** button.

   The Import Annotations from File window opens.



6. In the **Column Titles** box, do the following:
   a. If your file has annotation titles, select the **Annotation file has column titles** option.
   b. To indicate on which row the annotation titles appear, enter the number of the row in which the titles appear in the **Line of column titles**. For example, if the titles appear on row 2, enter the number 2.

If you do use column titles as annotation names, you can choose the name of the column in the file as the name of the annotation. If you do not use column titles as annotation names, you can either choose one of the standard columns or enter a custom name.

7. To set a “Systematic Name” column, locate the column you want to label “Systematic Name.”

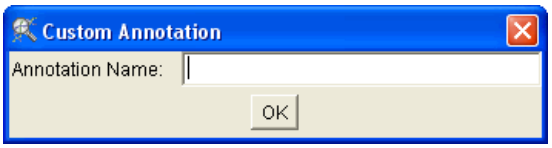
   Note: The file you import into GeneSpring does not need to have a Systematic Name column. If you don’t have a Systematic Name in your update file, you will not be able to load any new genes because new genes need a Systematic Name. You can, however, still update the annotation in the Master Table of Genes.

8. Select the **Systematic Name** option from the menu.

9. If a column in the annotations file is blank or you do not want to import the annotation column, choose the **Ignore** option in the menu.

10. If the annotation type is not listed and you do not want to use the title of the column in the update file as the name of the annotation, choose the **Custom** option in the menu.

    The Custom Annotation window opens.



11. Enter the name you want for the annotation column and click the **OK** button.

    The name appears as the column header.

12. In the **Recognizing Existing Genes** box, Choose the column you want to use to match the names by selecting the column name from the drop down menu

    When you want to update existing annotations (or add new genes and annotations), you do not have to match the Systematic Name from the update file to the Systematic Name in the Master Table of Genes.

    If you have an update file that contains a column of identifiers other than the Systematic Name, (for example, GenBank accession numbers or HUGO Gene symbols), you can use those identifiers to match the identifier in the Master Table of Genes. If you have a systematic name column in the file, it will be used automatically.

    You should first have completed the setting of the Annotation type before setting the Match name column, since the drop down menu uses the names of the columns. If the column in the update file does not have the same name as the annotation column in the Master Table of Genes, ensure that you assign the correct name to the column. You can either use one of the standard column names, or choose a custom column name that matches the column in the Master Table of Genes.

13. Click the **OK** button.

    GeneSpring reviews the edits you made. Depending on what you have done, the following actions can occur:

    a. GeneSpring will review the file and report any multiple matches of identifiers, since it is possible that one or more identifiers in the update file match one or more identifiers in the Master Table of Genes. If one identifier from the update file matches multiple identifiers in the Master Table of Genes, you will be asked what you want to do for the annotation columns.Choose the most appropriate option.

       - Ignore duplicates, update only unique entries
       - Update annotations for all entries

       If one identifier from the update file is repeated several times in the update file, you will be asked what you want to do. There is also an indication whether or not the annotations that are being loaded are identical or not. Choose the most appropriate option:

       - Ignore duplicates, use first entry
       - Ignore duplicates, use last entry
       - Ignore duplicates, import only unique entries
       - Append the annotations for all the entries

    b. If the Duplicate Identifiers window opens, GeneSpring has detected a problem with the annotation. Go to "Fixing Duplicate Identifiers" on page 6-49 to resolve the problem.

    c. If the Name Problems window opens, if GeneSpring has detected a problem with long gene names. Go to "Fixing Long Gene Name Problems" on page 6-51 to resolve the problem.

    If the edits are successful, the Import Annotations Summary window opens. Go to "Completing the Import Genes and Annotations Process" on page 6-52 for more information.

Fixing Duplicate Identifiers

This section explains how to handle duplicate identifiers when importing annotations from a file.

Fix Duplicate Identifiers Window

The Duplicate Identifiers window (Figure 6-11) appears if there are duplicate gene entries in the annotations file, or if the gene identifiers in the file match multiple existing gene.

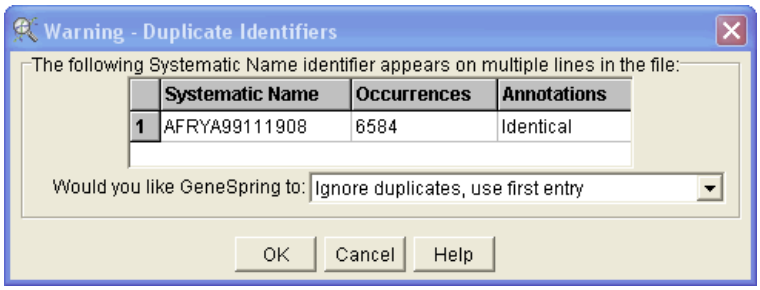


Figure 6-11 Import Annotations from File: Duplicate Identifiers Window

This window contains the following elements:

- **Gene identifier**—Displays the ID for each repeated entry. The title of this column is the name of the column chosen to match existing genes.
- **Occurrences**—Displays the number of times the systematic name occurs.
- **Annotations**—Compares the annotations you have selected to import and lets you know if all instances of the same gene name have the same annotations. The entries in this column are either Identical or Not Identical.
- **Systematic Name**—Displays the systematic name for each repeated entry.
- **Second column**—Displays the gene identifier used in the annotations file that is imported. The title shows the name of that annotation column.

Fixing Duplicate Identifiers

To fix duplicate identifier entries:

1. In the **Following No. of Identifiers appear on multiple lines in the file** box, select one of the following options in the **Would You Like GeneSpring To** menu:
   - Ignore Duplicates, use the first entry.
   - Ignore Duplicates, use the last entry.
   - Ignore Duplicates, import only unique entries.
   - Append annotations for all of the entries.

     Selecting this option appends annotations to all entries in the window.
2. In the **Following No. of Identifiers appear in the genome multiple times** box, select one of the following options in the **Would You Like GeneSpring To** menu:
   - Ignore Duplicates, update only unique entries.
   - Update annotations for all entries.
3. Click the **OK** button.

   If the edits are successful, the Import Annotations Summary window opens. Go to "Completing the Import Genes and Annotations Process" on page 6-52 for more information.

Fixing Long Gene Name Problems

This section explains how to fix long gene name problems when importing annotations from a file.

Name Problems window

The Name Problems window (Figure 6-12) indicates that GeneSpring has detected a problem with long gene names while importing annotations from a file. The limit for name in the Systematic Names, Common names, GenBank accession numbers, and Synonyms columns must be shorter than 256 characters. Long gene names are not allowed and must be resolved before you can proceed.

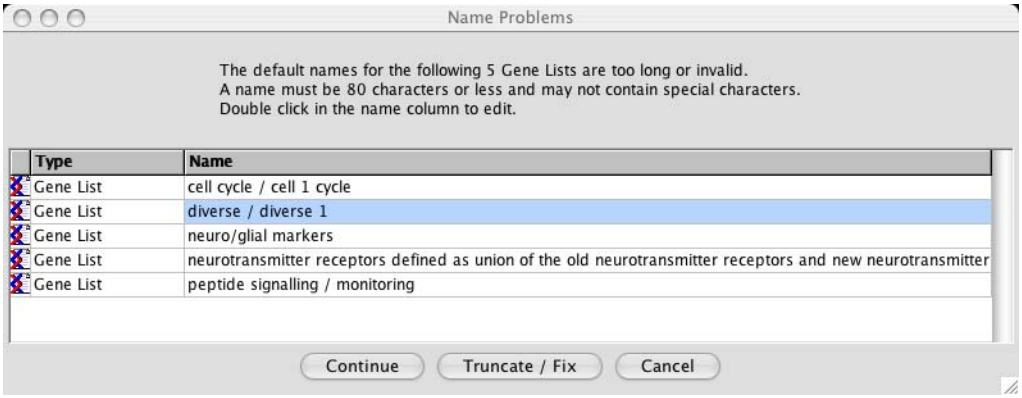

Name Problems
The default names for the following 5 Gene Lists are too long or invalid.
A name must be 80 characters or less and may not contain special characters.
Double click in the name column to edit.
Type
Name
Gene List
cell cycle / cell 1 cycle
Gene List
diverse / diverse 1
Gene List
neuro/glial markers
Gene List
neurotransmitter receptors defined as union of the old neurotransmitter receptors and new neurotransmitter
Gene List
peptide signalling / monitoring
Continue
Truncate / Fix
Cancel

Figure 6-12 Import Annotations from File: Name Problems Window

This window contains the following elements:

- **Table**—Displays the results that have name problems.
- **Continue**—Disabled until the problems are fixed.
- **Truncate/Fix**—Truncates or fixes all of the problem entries based on their allowable size and characters.
- **Cancel**—Cancels the operation and closes the window.

To fix long names or invalid characters:

1. Do one of the following:
   - Manually fix all of the problem entries by editing them.
   - Automatically fix all of the problem entries by clicking the **Truncate/Fix** button.
2. Click the **Continue** button to fix the problem.

   The Import Genomic Sequence window opens.

   If the edits are successful, the Import Annotations Summary window opens. Go to "Completing the Import Genes and Annotations Process" on page 6-52 for more information.

Completing the Import Genes and Annotations Process

This section explains how to complete the annotation import process.

Using the Annotations Summary Window

After GeneSpring scans the file and verifies that there will be no problems importing it, the Annotations Summary window (Figure 6-13) opens.

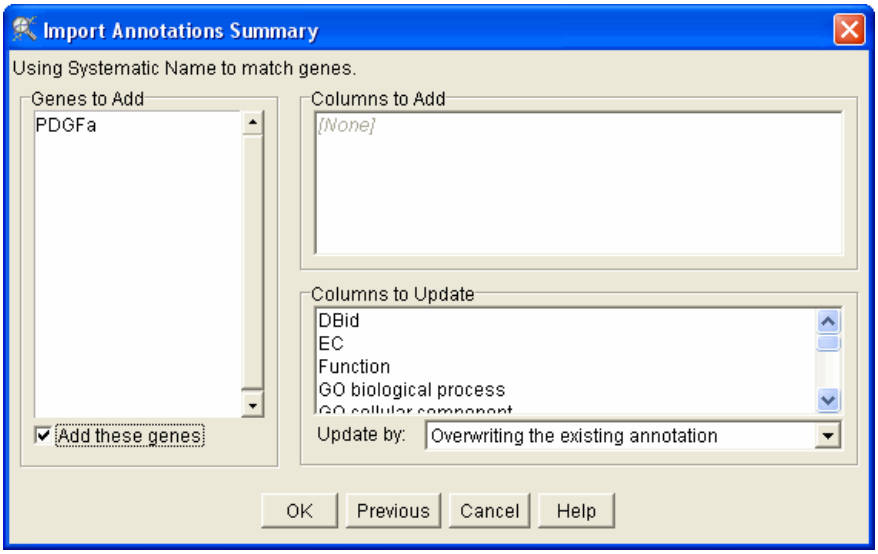


Figure 6-13 Annotations Summary Window

This window contains the following elements:

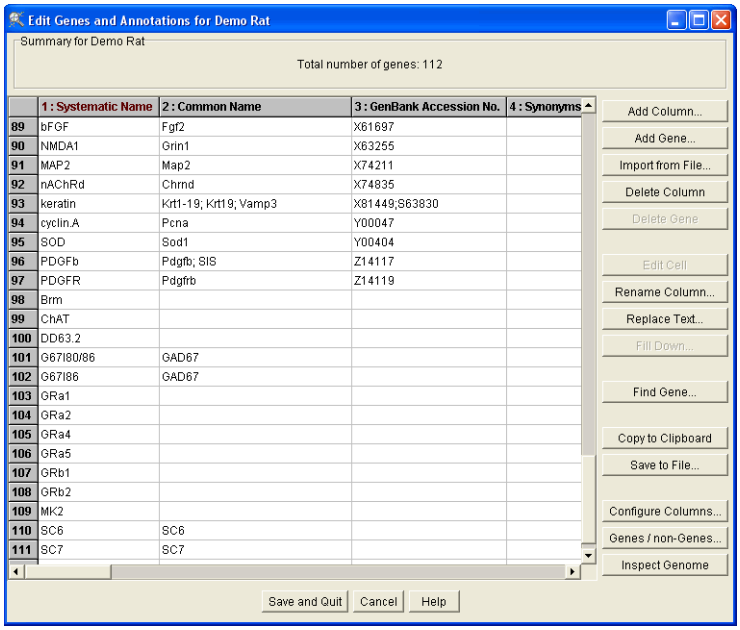
- **Genes to Add box**—Lists all of the genes that will be added due to the importation of this new annotation file. This box is not available at all if you are importing genes based on a gene identifier that is not the Systematic Name column.
- **Columns to Add box**—Lists all of the new annotations that will be added. This box is disabled if there were no columns selected that didn't already exist in the Master Table of Genes.
- **Columns to Update box**—Lists all of the annotations that will be updated by importing annotations from the external file. This box is disabled if there are no columns selected that already exist in the Master Table of Genes.
- **Update by box**—Determines what happens if there are annotation values already in the annotation column for one of the imported annotations.

Completing the Annotation Import Process

To complete the annotation import process:

1. In the **Update by box**, do the following:
   - To overwrite an existing annotation upon import, select the **Overwriting the existing annotation** option
   - To append an existing annotation upon import, select the **Appending to the existing annotation** option.
   - To fill in the blank cells upon import, select the **Filling in the blanks in the existing annotation** option.
2. Click the **OK** button.

   The annotations are updated and the Master Table of Gene Editor reappears.



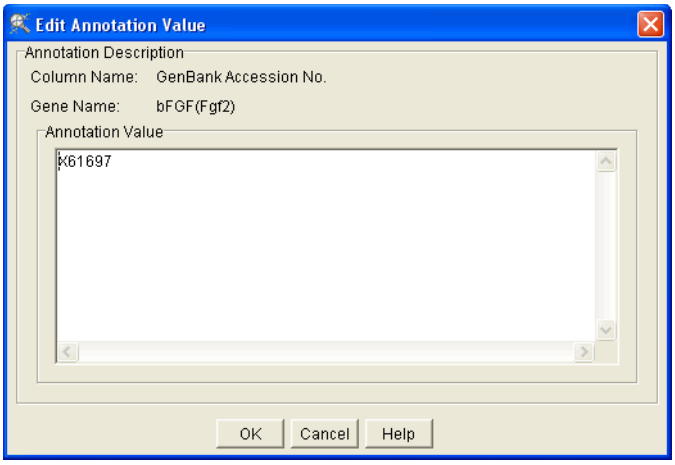
Editing Annotations

To edit annotations:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.
2. Select the cell you want to edit and click the **Edit Cell** button.

The Edit Annotation Value window opens.



3. In the **Annotation Value** box, enter the annotation value you want.

   Note: You can also cut and paste the annotation value using the menu.

4. Click the **OK** button.

Finding Genes

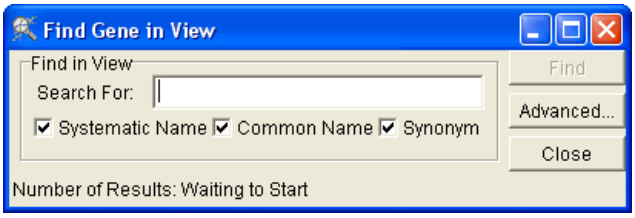
To find genes:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.

2. Click the **Find Gene** button.

   The Find Gene in View window opens.



3. To find a gene in the current view, do the following:

   a. In the **Search For** field, enter the name of the gene you want to find.

   b. To narrow the search, deselect the **Systematic Name**, **Common Name,** or **Synonym** options.

   c. Click the **Find** button.

      Results are highlighted in the table.

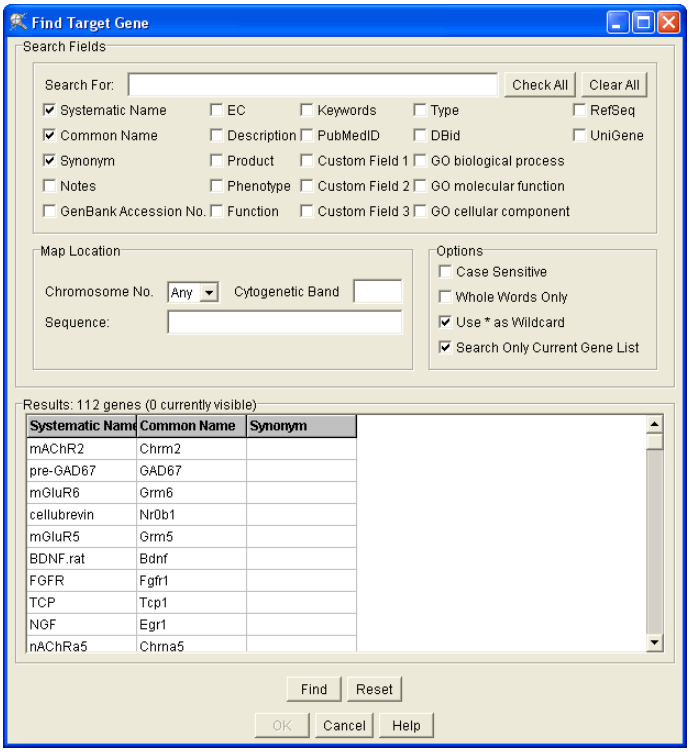
   d. Click the **Next** button to view additional matches.

e. When you are done, click the **Close** button.

4. To specify advanced search options, do the following:

   a. Click the **Advanced** button.

      The Find Target Gene window opens.



5. In the **Annotation** check boxes, do one of the following:

   - To search individual annotations, select the check box you want.
   - To search all annotations, click the **Check All** button.
   - To clear all selected annotations, click the **Clear All** button.

6. In the **Search For** field, enter the search terms you want:

   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

   White space in the Search For field behaves as an AND between words or phrases.

- Enter an asterisk '*' in this field to match any gene with an entry in any of the fields you specified.
- You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
- To perform an exact string match, use double quotes in your search term. For example, "expression profile".
- For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

7. To restrict your search, you can select any of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string "statistic", the search results will contain samples that contain the word "statistic" in the specified fields, but ignore those containing "statistic" as part of a larger word, such as "statistician".
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, "ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
   - **Search Only Current Gene List**—Searches the current gene list only.
8. Click the **Find** button.

   A list of the genes that match the search criteria is displayed in a table at the bottom of the window. The field that matched the search criteria is colored red.

9. To learn more about one or more genes, do one of the following:
   - Select the gene of interest in the appropriate row.
   - To select more than one gene, hold down the **Ctrl** key while clicking multiple rows.
10. Once you have highlighted the gene(s) of interest, do the following:
    - To select the genes in the Genome Browser, click the **Select** button.
    - To select all of the genes identified by the search, click **Select All** button.
    - To select and zoom-in on a gene in the Genome Browser, click **Zoom & Select** button.
    - To display the Gene Inspector window for the selected gene, click **Inspect** button.
    - To make a gene list from all of the genes identified by the search, click **Make Gene List** button.
11. Click the **Find** button.
12. When you are done searching, click the **OK** button.

    Results appear at the bottom of the window.
13. When you are done, click the **Close** button.

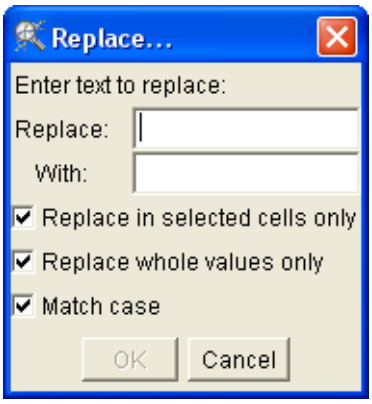
Finding and Replacing Text

To find and replace text:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.
2. Click the **Replace Text** button.

   The Replace window opens.



3. In the **Replace** field, enter the text you want to replace.
4. In the **With** field, enter the new text.

5. Select one or more of the following options:
   - Replace in selected cells only.
   - Replace whole values only.
   - Match case.
6. Click the **OK** button.

Copying Data to the Clipboard

This procedure copies the entire Master Table of Genes to the clipboard, regardless of what is selected in the table.

To copy data to the Clipboard:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.
2. Click the **Copy to Clipboard** button.
3. Paste the data in the target application.

Saving Data to a File

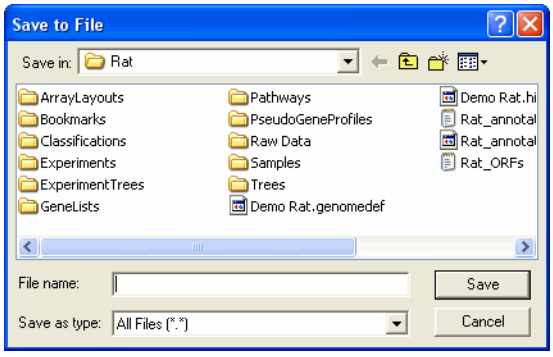
The following procedure saves the entire Master Table of Genes regardless of what row, column, or cell is selected in the table.

To save data to a file:

1. Open the **Master Table of Genes Editor**.

   Go to "Opening the Master Table of Genes Editor" on page 6-42 for instructions.
2. Click the **Save to File** button.

   The Save to File window opens.



3. Navigate to the directory you want.
4. Click the **Save** button.

Adding Other Genomic Elements

This section explains how to add other genomic element types to the Master Table of Genes. It covers the following topics:

- Define Other Genomic Elements Window
- Opening the Define Other Genomic Elements Window
- Filtering on Gene Lists
- Filtering on Systematic Name, Common Name, Synonym, or GenBank ID
- Filtering on Annotations
- Adding Other Genomic Elements

Define Other Genomic Elements Window

The Define Other Genomic Elements window (Figure 6-14) lets you create a new gene list based on a variety of selection criteria, add other genomic elements to the list of other genomic elements, or remove other genomic elements from the lists. Other genomic elements include telomeres, rRNA genes, and CpG islands that are usually not directly transcribed into mRNA.

Genes that are added to the Define Other Genomic Elements window will be treated as other genomic elements. Genes that are removed from this window will be treated as genes.

Note: When you modify the genomic element types, you'll need to restart GeneSpring to enable the changes.

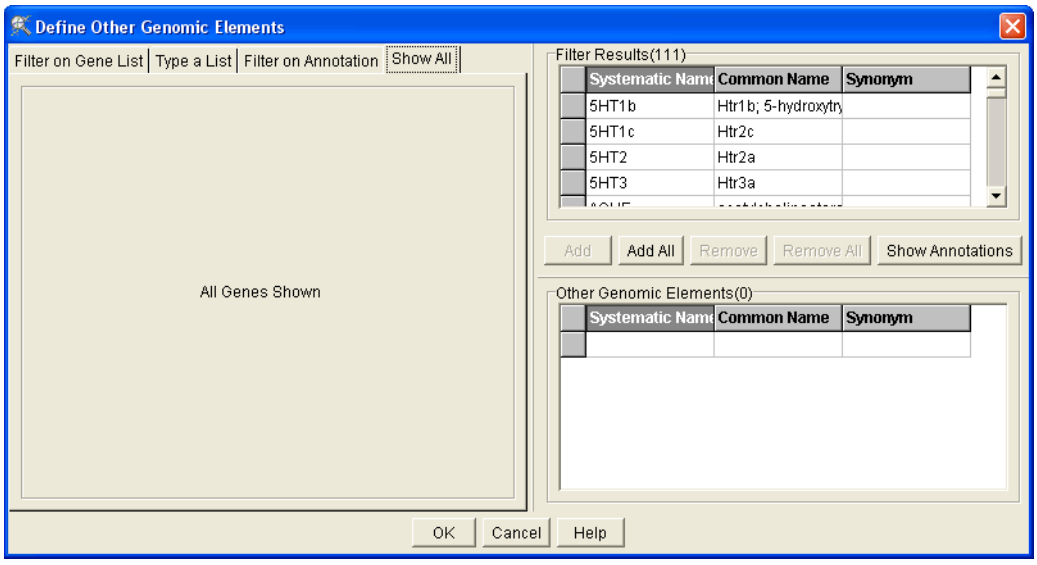


Figure 6-14 Other Genomic Elements Window: Show All Tab

Filtering Options

The left side of the window contains tabs for each filtering method. Click a tab to view options for that method. The methods are:

- **Show All**—Display all available genes without applying a filter.
- **Filter on Annotation**—Display genes based on a specified annotation.
- **Type a List**—Manually enter a list of genes.
- **Filter on Gene List**—Display genes from a selected gene list.

Filter on Gene List Tab

To filter on gene list, use the Filter on Gene List tab (Figure 6-15) to locate the desired gene list and select it in the list. All genes in that list appear in the Filter Results table.

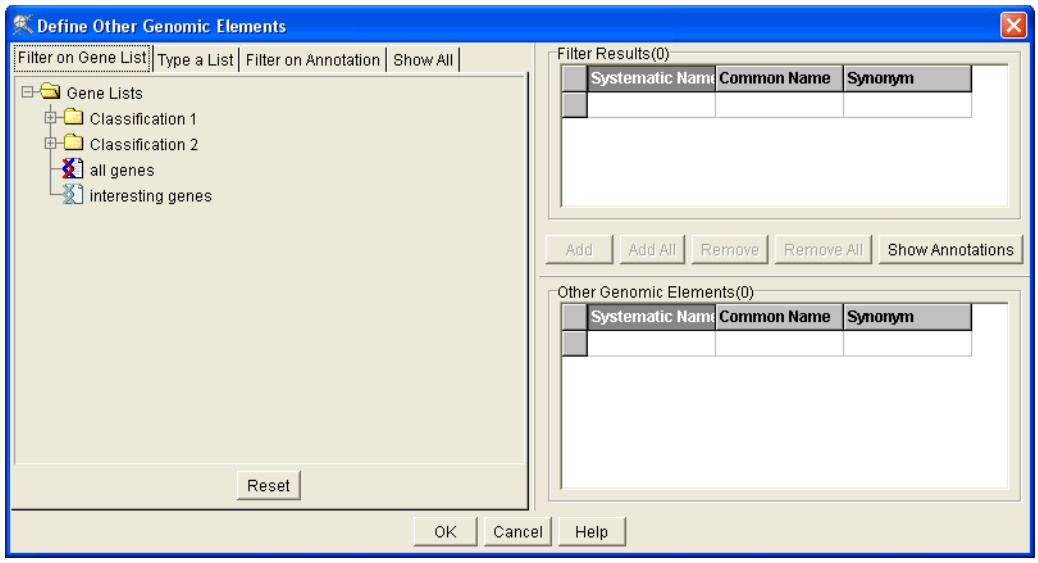


Figure 6-15 Other Genomic Elements Window: Filter on Gene List Tab

Type a List Tab

The Type a List tab (Figure 6-16) allows you to manually enter a list of genes. To enter genes, simply click in the gene list box, type a gene's Common Name, Systematic Name, Synonym, or GenBank Accession Number and press Enter. You can also use copy and paste to enter one or more genes to this list.

If the gene is found, it appears in the Filter Results table. If the gene is not found, it is colored in red. To clear your entries, click Reset. This removes all entered genes from both the list of genes you have typed in and the Filter Results table.

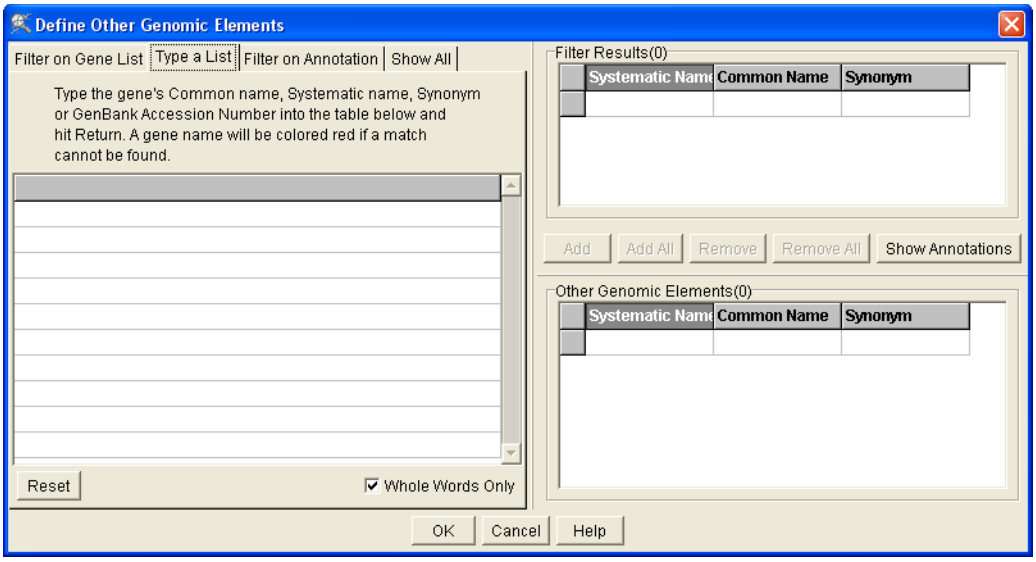


Figure 6-16 Other Genomic Elements Window: Type a List Tab

Filter on Annotation Tab

The Filter on Annotation tab (Figure 6-17) allows you to filter genes based on text or values in a specified annotation.

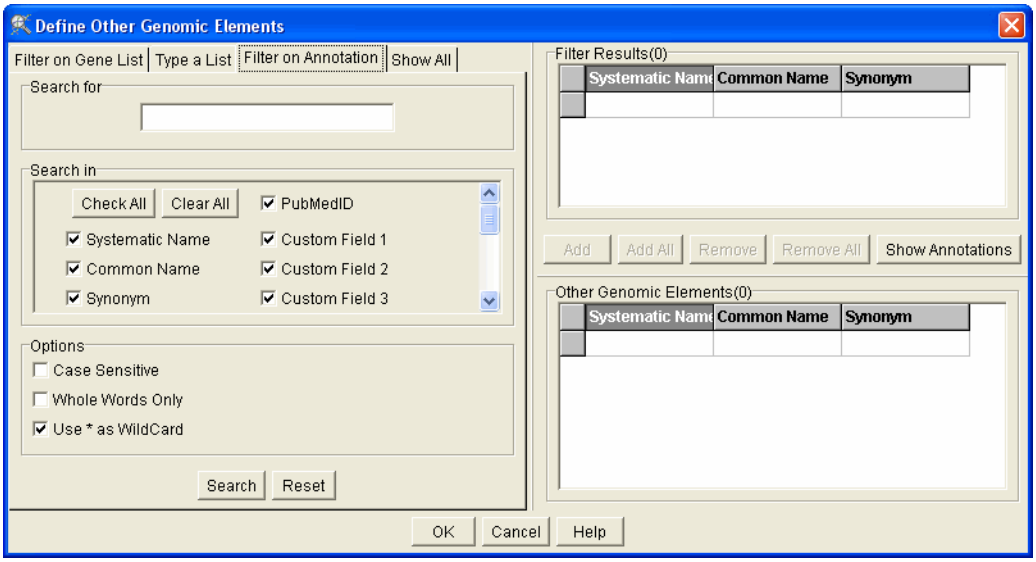


Figure 6-17 Other Genomic Elements Window: Filter on Annotation Tab

Filter Results Table and New Gene List Table

The right section of the window contains two tables: Filter Results and New Gene List. The upper table contains all of the genes resulting from the current filtering method. The lower table contains the genes you have selected to add to your list. Between the two tables are five buttons.

Buttons

- **Add**—Adds a selected gene in the Filter Results table to the New Gene List table.
- **Add All**—Adds all genes in the Filter Results table to the New Gene List table.
- **Remove**—Removes a selected gene from the New Gene List table.
- **Remove All**—Removes all genes from the New Gene List table.
- **Show/Hide Annotations**—Shows or hides annotations from a gene list.

Opening the Define Other Genomic Elements Window

To open the Define Other Genomic Elements window:

1. Do one of the following:
   - Select **Annotations > Edit Genes and Annotations**.

     Note: You cannot open the Master Table of Genes Editor when a Homology Table window (for any genome) is open or when the GeneSpider is running.
   - Open the **Genome Inspector**, click the **Genome Properties** tab, and then click the **Edit Genes and Annotations** button.

     Go to "Opening the Genome Inspector" on page 6-20 for instructions.

     The Master Table of Genes Editor opens.
2. Click the **Genes/Non Genes** button.

   The Define Other Genomic Elements window opens.

Filtering on Gene Lists

To filter on a gene list:

1. Open the **Other Genomic Elements** window.

   Go to "Opening the Define Other Genomic Elements Window" on page 6-62 for complete information.
2. Click the **Filter on Gene List** tab.
3. In the **Gene Lists** folder, select the list you want.

   Genes matching the specified search parameters appear in the Filter Results table.
4. To add the other genomic elements, go to "Adding Other Genomic Elements" on page 6-65.

Filtering on Systematic Name, Common Name, Synonym, or GenBank ID

To filter on Systematic Name, Common Name, Synonym, or GenBank ID:

1. Open the **Other Genomic Elements** window.

   Go to "Opening the Define Other Genomic Elements Window" on page 6-62 for instructions.

2. Click the **Type a List** tab.
3. Enter the gene's Systematic Name, Common Name, Synonym, or GenBank ID.
4. To search by whole word, select the **Whole Word** check box.
5. Click the **OK** button.

   Genes matching the specified search parameters appear in the Filter Results table.

6. To add the other genomic elements, go to "Adding Other Genomic Elements" on page 6-65.

Filtering on Annotations

To filter on annotations:

1. Open the **Other Genomic Elements** window.

   Go to "Opening the Define Other Genomic Elements Window" on page 6-62 for complete information.

2. Click the **Filter on Annotations** tab.
3. In the **Search for** field, enter the search terms you want:

   - Create a Boolean search by entering multiple terms separated by the words "and", "or", or "not" in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.
   - Enter an asterisk '*' in this field to match any gene with an entry in any of the fields you specified.
   - You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
   - To perform an exact string match, use double quotes in your search term. For example, "expression profile".

- For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

4. In the **Search in** field, do one of the following:
   - To search individual annotations, select the check box you want.
   - To search all annotations, click the **Check All** button.
   - To clear all selected annotations, click the **Clear All** button.
5. To restrict your search, you can select any of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
6. Click the **Search** button.

   Genes matching the specified search parameters appear in the Filter Results table.
7. To add the other genomic elements, go to “Adding Other Genomic Elements” on page 6-65.

Adding Other Genomic Elements

To add other genomic elements:

1. In the **Other Genomic Elements: Filter Results** table, do the following:
   - To add a selected gene in the Filter Results table to the New Gene List table, click the **Add** button.
   - To **Add All** genes in the Filter Results table to the New Gene List table, click the Add All button.
2. Click the **OK** button.
3. Restart GeneSpring to enable the changes.

7

Finding, Selecting, and Inspecting Data

This chapter explains the different options that are available to find, select, inspect, and manage genes or associated data objects. The following topics are covered:

- Finding Genes
- Selecting Genes
- Searching for Data Locally or Remotely
- Using Inspectors
- Working with Projects

Finding Genes

This section describes the tools that are available to find genes and other data objects in GeneSpring. It covers the following topics:

- Using the Genome Browser
- Finding Genes
- Using the Advanced Find Gene Function

Using the Genome Browser

The large panel in the center of the GeneSpring window is the Genome Browser (Figure 7-1), which graphically displays information about the genes in the selected gene list. You can use the Genome Browser to simply browse for genes using different views, like the Graph view or the Physical Position view. However, the Genome Browser often presents so much information that individual genes and gene names are not visible. This section explains how to use the Genome Browser to look at genes more closely.

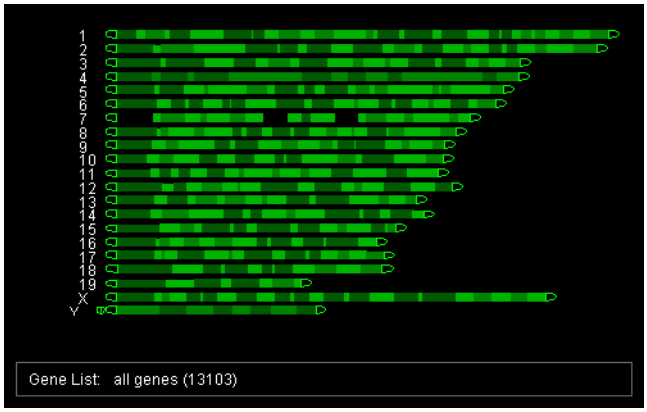


Figure 7-1 Genome Browser: Physical Position View

Zooming

You can enlarge or reduce a region of a window by zooming. Zooming in increases magnification step-by-step (Figure 7-2). Zooming out decreases magnification step-by-step.

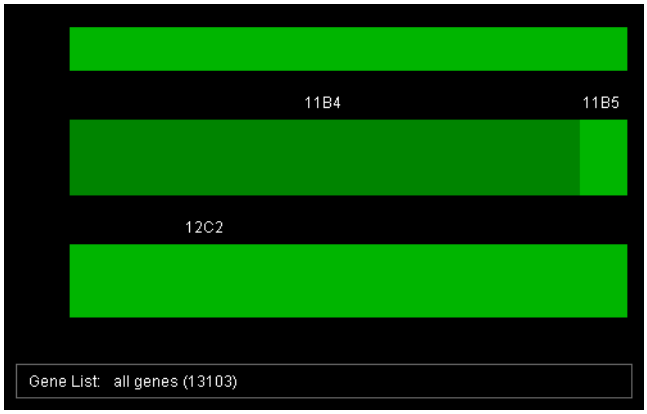


Figure 7-2 Genome Browser: After Zooming In

To zoom in the Genome Browser:

- Do one of the following:
  - Click Ctrl+].
  - Click and drag a rectangle across the region to enlarge, and then release the mouse button. Repeat this step until you achieve the level of magnification you want.

To zoom out the Genome Browser:

- Do one of the following:
  - Type Ctrl+[.
  - Click **Zoom Out**.

  This will decrease the magnification two times.

To return directly to the unmagnified state:

- Do one of the following:
  - Select **View > Zoom Fully Out** option.
  - Click **Zoom Fully Out**.
  - Type **Ctrl + Home**.

Panning

In the magnified view, you can pan in any direction.

To pan the Genome Browser:

- Do one of the following:
  - Press the **Up**, **Down**, **Left**, or **Right** arrow key to move in the direction you want.
  - Press the **Page Up** or **Page Down** keys to move up or down the window.

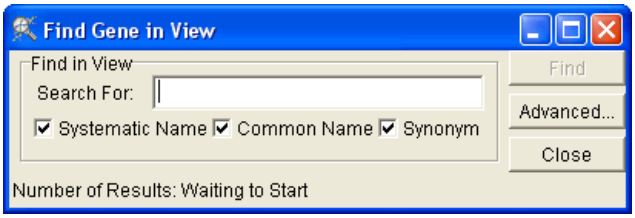
Finding Genes

The Find Gene function lets you quickly find one or more genes. This tool is especially useful when there are too many genes in the Genome Browser to easily identify individual genes.

To find genes:

1. Select **Edit > Find Gene**.

   The Find Gene in View window opens.



2. In the **Search For** field, enter the name of the gene you want to find.
3. To narrow the search, deselect the **Systematic Name**, **Common Name,** or **Synonym** options.

4. Do one of the following:
   - Click the **Find** button.
   - Press the **Enter** key.

   In some views, the Genome Browser zooms in on the gene. This gene is automatically selected. If more than one gene is found that matches your search criteria, the total number of genes found are listed at the bottom of the Find Gene Window along with the number of genes found that are visible in the current gene list.
5. If additional genes are found, click **Find Next** to show the next gene that matches your search criteria.

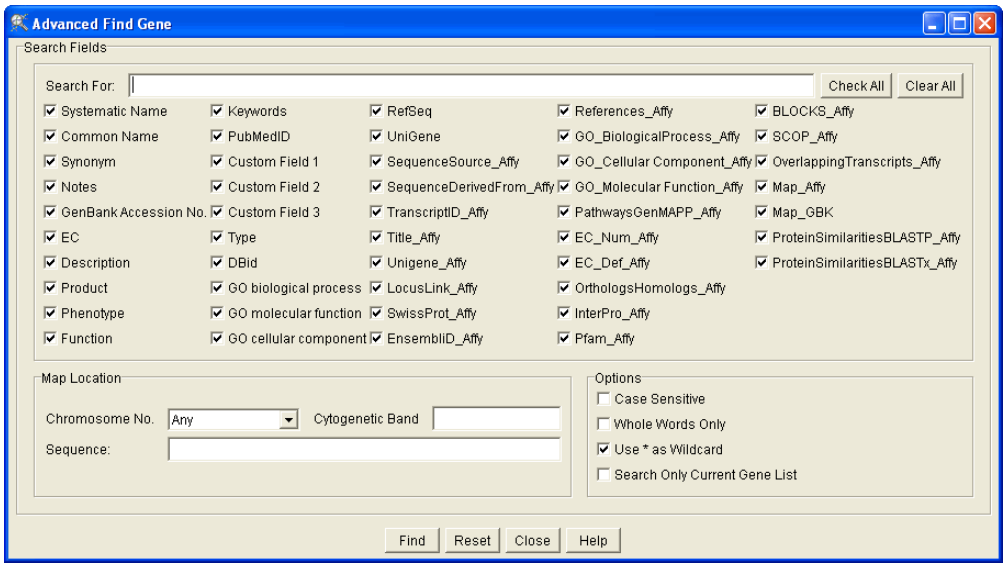
Using the Advanced Find Gene Function

The Advanced Find Gene function provides the same search criteria as the Find Gene function, but also provides more advanced criteria to expand or narrow your search.

To use the Advanced Find Gene function:

1. Do one of the following:
   - Select **Edit > Find Gene** and click the **Advanced** button.
   - Select **Edit > Advanced Find Gene**.

   The Advanced Find Genes window opens.



2. In the **Annotation** check boxes, do one of the following:
   - To search individual annotations, select the check box you want.
   - To search all annotations, click the **Check All** button.
   - To clear all selected annotations, click the **Clear All** button.
3. In the **Search For** field, enter the search terms you want:

- Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
  - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
  - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
  - **NOT**—Matches any gene containing the first term, but not containing the second.

  White space in the Search For field behaves as an AND between words or phrases.
- Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified in step 2.
- You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
- To perform an exact string match, use double quotes in your search term. For example, “expression profile”.
- For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

4. To restrict your search, you can select any of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.

- Words with an * in them or next to them will not have the Whole Words Only Search applied.
- If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.

- **Search Only Current Gene List**—Searches the current gene list only.

5. Click the **Find** button.

   A list of the genes that match the search criteria is displayed in a table at the bottom of the window. The field that matched the search criteria is colored red.

6. To learn more about one or more genes, do one of the following:
   - Select the gene of interest in the appropriate row.
   - To select more than one gene, hold down the **Ctrl** key while clicking multiple rows.
7. Once you have highlighted the gene(s) of interest, do the following:
   - To select the genes in the Genome Browser, click the **Select** button.
   - To select all of the genes identified by the search, click **Select All** button.
   - To select and zoom in on a gene in the Genome Browser, click the **Zoom & Select** button.
   - To display the Gene Inspector window for the selected gene, click the **Inspect** button.
   - To make a gene list from all of the genes identified by the search, click the **Make Gene List** button.
8. When you are done, click the **Close** button.

Searching for Data Locally or Remotely

The Search window lets you search for genes, experiments, samples, or other data objects that are stored locally on GeneSpring or remotely on Signet. From this window, you can search for either genes or associated data objects using the following criteria:

- Search for data by keyword
- Search for data objects by project
- Search for data using advanced search criteria
- Search for an individual gene

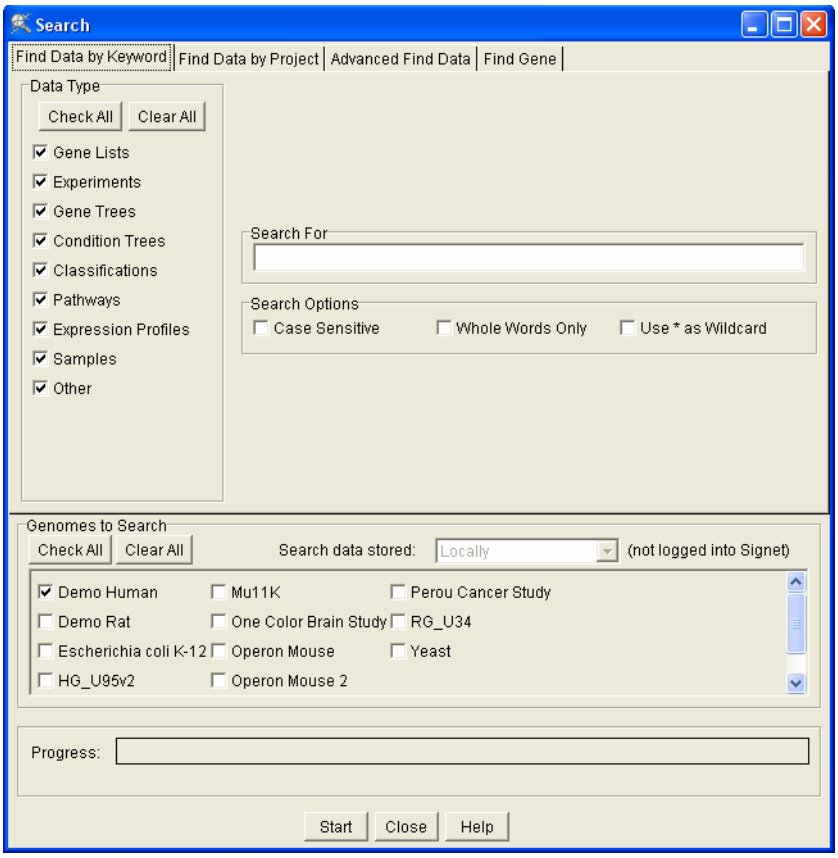
To search Signet, you must be logged in to a remote execution server. Moreover, Signet searches will only return data you have permission to view. If you are logged into more than one Signet server, a dialog opens that prompts you to select the Signet server on which to search. You cannot search multiple Signet servers at the same time.

Searching for Data by Keyword

To search for data by keyword:

1. Select **Edit > Search**.

   The Search window opens and the Find Data by Keyword tab appears.



2. In the **Data Type** box, select the Navigation folders in which to search.
3. In the **Search For** field, enter the search terms you want.
   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.
   - Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified in step 2.

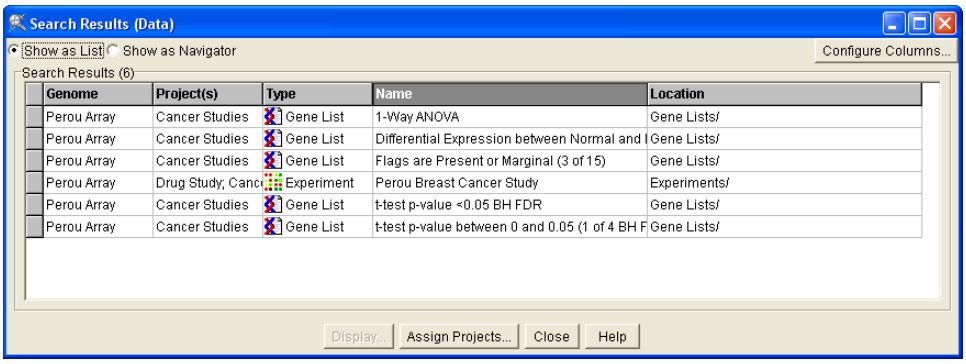
- You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
- To perform an exact string match, use double quotes in your search term. For example, “expression profile”.
- For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

4. In the **Search Options** box, restrict your search by choosing one of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you In the **Genomes to Search** box, select the genomes in which to search.

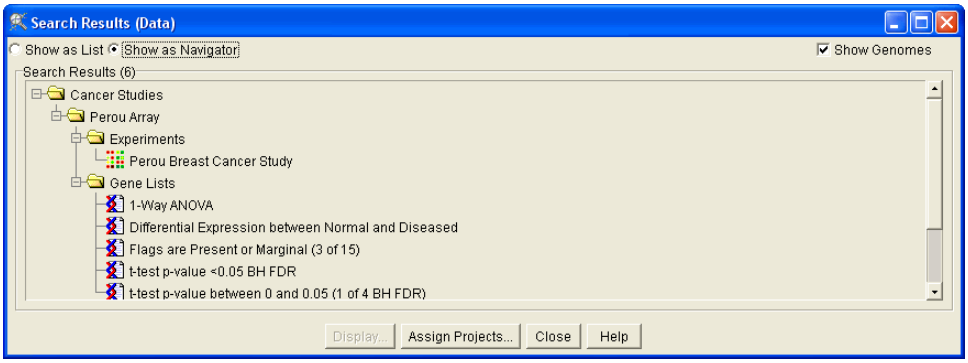
   The more genome names you check, the longer the search process will take.
5. In the **Search data stored** menu, specify whether to conduct the search locally, on Signet, or on both.

   Note: You must be logged in to Signet to perform a remote search.
6. Click the **Start** button.

   When the search is complete, a list of the data objects that match the search criteria appears in the Search Results window opens in the **Show as List** view.



7. To show the results in the Navigator, select the **Show as Navigator** option.



8. To assign the results to a project, go to "Working with Projects" on page 7-40 for more information.

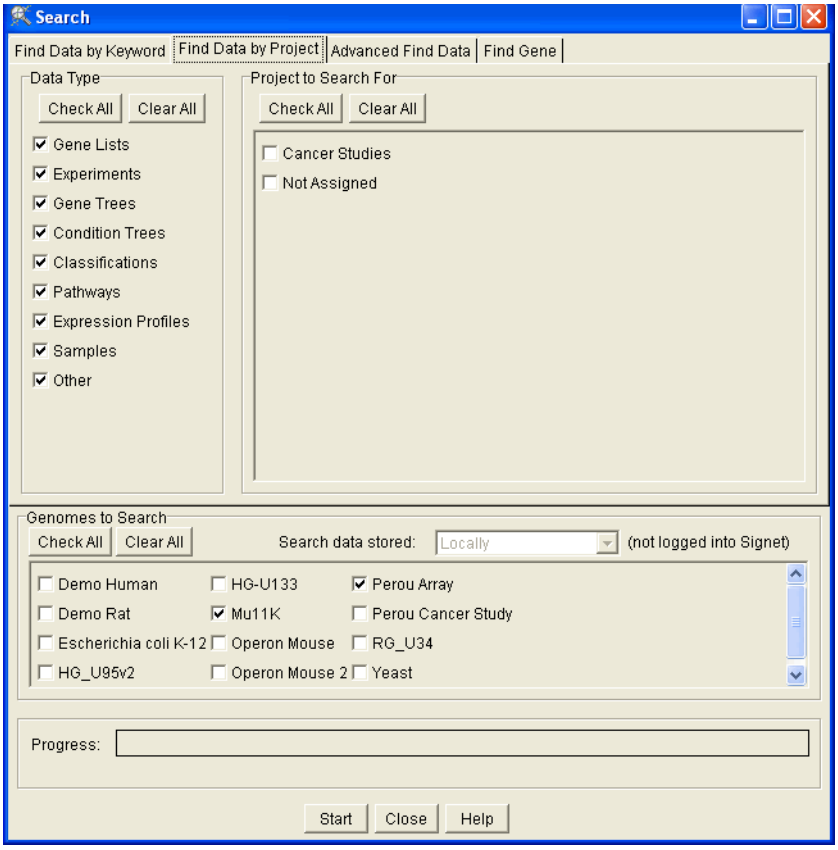
Searching for Data by Project

To search for data by project:

1. Select **Edit > Search**.

   The Search window opens.

2. Click the **Find Data by Project** tab.



3. In the **Data Type** box, select the Navigation folders in which to search.
4. In the **Project to search for** box, select the projects in which to search.
5. In the **Genomes to Search** box, select the genomes in which to search.

   The more genome names you check, the longer the search process will take.
6. In the **Search data stored** menu, specify whether to conduct the search locally, on Signet, or on both.

   Note: You must be logged in to Signet to perform a remote search.
7. Click the **Start** button.

   When the search is complete, a list of the data objects that match the search criteria appears in the Search Results window opens in the **Show as List** view.
8. To show the results in the Navigator, select the **Show as Navigator** option.
9. To assign the results to a project, go to "Working with Projects" on page 7-40 for more information.

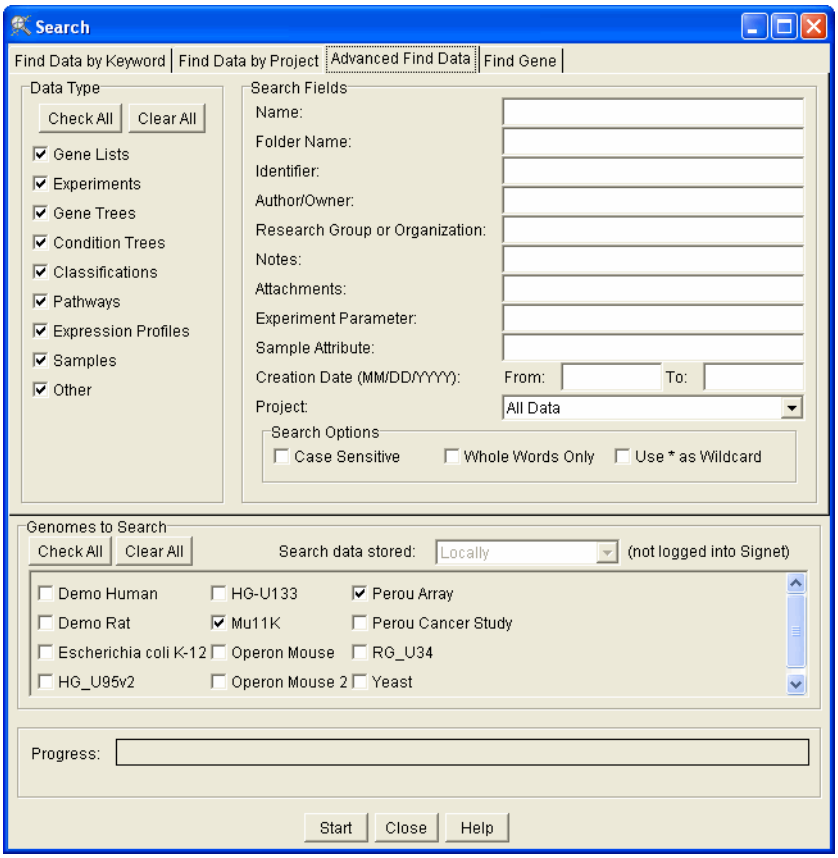
Searching for Data Using Advanced Criteria

To search for data using advanced criteria:

1. Select **Edit > Search**.

   The Search window opens.

2. Click the **Advanced Find Data** tab.



3. In the **Data Type** box, select the **Navigation** folders in which to search.
4. In the **Search Options** box, restrict your search by choosing one of the following options:

   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.

- **Search Only Current Gene List**—Searches the current gene list only.

5. In the **Search Fields** box, enter the information you want in the fields provided.
6. In the **Genomes to Search** box, select the genomes in which to search.

   The more genome names you check, the longer the search process will take.
7. In the **Search data stored** menu, specify whether to conduct the search locally, on Signet, or on both.

   Note: You must be logged in to Signet to perform a remote search.
8. Click the **Start** button.

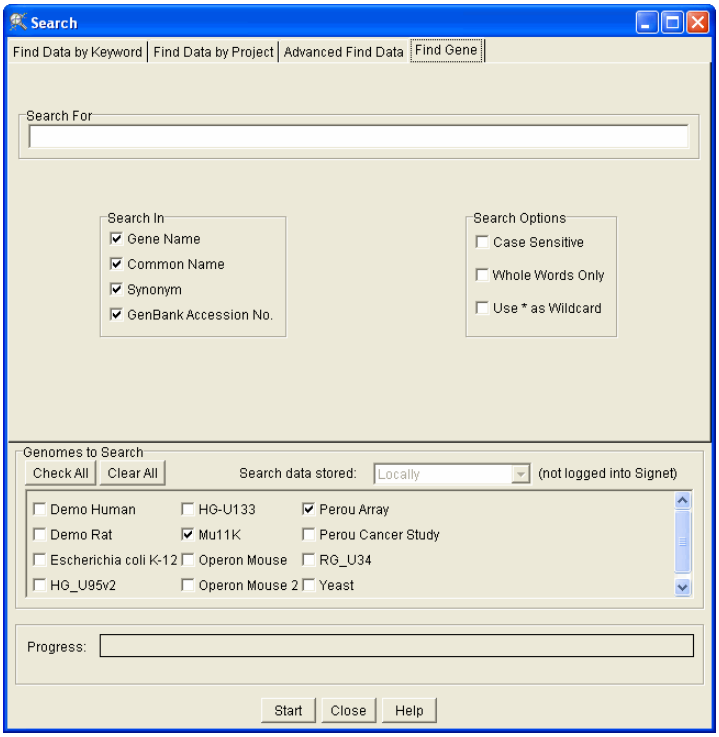
   When the search is complete, a list of the data objects that match the search criteria appears in the Search Results window opens in the **Show as List** view.
9. To show the results in the Navigator, select the **Show as Navigator** option.
10. To assign the results to a project, go to "Working with Projects" on page 7-40 for more information.

Searching for a Gene

To search for a gene:

1. Select **Edit > Search**.

   The Search window opens.
2. Click the **Find Gene** tab.



3. In the **Search For** field, enter the search terms you want:

   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.

     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.

   - To perform an exact string match, use double quotes in your search term. For example, “expression profile”.

4. In the **Search Options** box, restrict your search by choosing one of the following options:

   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.

5. In the **Search In** box, select the annotations you want to search.
6. In the **Genomes to Search** box, select the genomes in which to search.

   The more genome names you check, the longer the search process will take.

7. In the **Search data stored** menu, specify whether to conduct the search locally, on Signet, or on both.

   Note: You must be logged in to Signet to perform a remote search.

8. Click the **Start** button.

   A list of the genes that match the search criteria appears in the Search Results (Genes) window. The field that matched the search criteria is colored red.

Selecting Genes

During the course of your research, you might find the need to select a gene or group of genes with which to work.

Selecting a Single Gene

To select a single gene in the Genome Browser:

1. Click once on any line or square representing a gene.

   The name of the selected gene appears in the legend at the bottom of the Genome Browser.

2. To inspect the selected gene, do one of the following:

   - Double-click a gene.

     This action works on genes represented graphically in the Genome Browser and on gene names found in lists.

   - Select a gene and press **Ctrl+I**.

   The Gene Inspector window opens. Go to "Using the Gene Inspector" on page 7-16 for more information.

   Tip: It is much easier to select a gene in the Genome Browser if you zoom in on it.

Selecting Multiple Genes

To select multiple genes in the Genome Browser:

1. Do one of the following:

   - Single-click on any line or square representing a gene, hold down Shift, and then select more lines or blocks to add more genes.

     Note: Clicking a selected gene while holding Shift deselects that particular gene.

   - Hold down Shift and drag your mouse across genes you want to select.

     A rectangle appears as you drag. When you release the mouse, the genes inside the rectangle are selected. When several genes are selected, the number of genes selected appears in the legend at the bottom of the Genome Browser.

2. To make a list of the selected genes, right-click in the Genome Browser, and select the **Make List from Selected Genes** command.

   If some selected genes are not in the current displayed gene list, the legend displays the message "*x* genes selected, *y* genes not in list" where *x* is the total number of selected genes and *y* is the number of selected genes not in the current gene list.

3. Click anywhere in the Genome Browser to clear the selected genes.

Learning More about Genes of Interest

Once you have found genes using the Find Genes or Advanced Find Genes functions, you can use several tools to obtain more information about them.

To learn more about genes of interest:

1. Do one of the following:
   - Select the gene of interest.
   - To select more than one gene, hold down the Shift key while clicking multiple rows.
2. Once you have highlighted the gene(s) of interest, do the following:
   - To select the genes in the Genome Browser, click the **Select** button.
   - To select all of the genes identified by the search, click **Select All** button.
   - To select and zoom-in on a gene in the Genome Browser, click the **Zoom & Select** button.
   - To display the Gene Inspector window for the selected gene, click the **Inspect** button.
   - To make a gene list from all of the genes identified by the search, click the **Make Gene List** button.

Using Inspectors

The GeneSpring Inspectors let you view the current settings and available details of any gene, condition, classification or experiment.The following Inspectors are described:

- Gene Inspector
- Sample Inspector
- Experiment Inspector
- Condition Inspector
- Gene List Inspector
- Classification Inspector

GeneSpring also provides a Script Inspector and an External Program Inspector. Go to Chapter 16, “Using Scripts, External Programs, and Plugins” for more information.

Using the Gene Inspector

The Gene Inspector lets you look at all the data associated with a particular gene, see the lists that include your gene, make correlations, and link directly to Internet databases.

Opening the Gene Inspector

To open the Gene Inspector:

- Do one of the following:
  - Double-click a gene.
  - Press **Ctrl+I** (when one or more genes are selected).

Gene Inspector Window

The upper-left corner of the Gene Inspector window (Figure 7-3) is called the Gene Identification section. It displays information about annotations. The table in the upper right corner displays the normalized, control, and raw values, as well as the t-test p-value and flag for each condition in the current experiment interpretations.

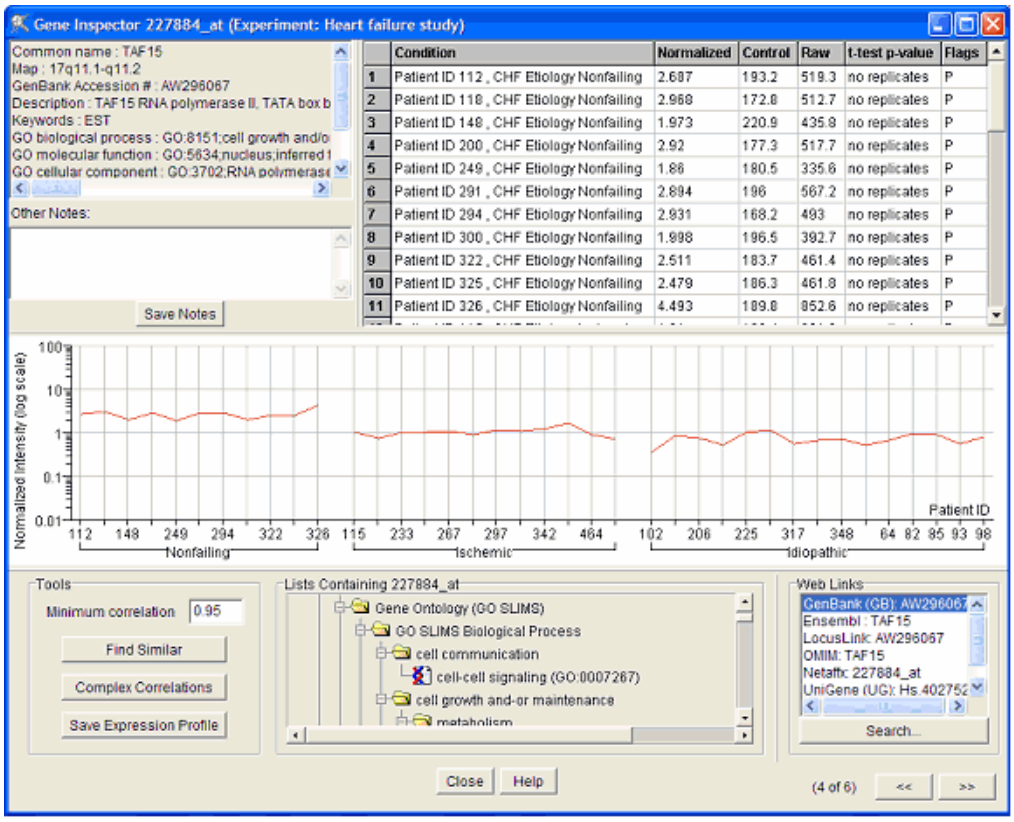

Gene Inspector 227884_at (Experiment: Heart failure study)
Common name : TAF15
Map : 17q11.1-q11.2
GenBank Accession # : AW296067
Description : TAF15 RNA polymerase II, TATA box b
Keywords : EST
GO biological process : GO:8151;cell growth and/o
GO molecular function : GO:5634;nucleus;inferred f
GO cellular component : GO:3702;RNA polymeras
Other Notes:
Save Notes
Condition | Normalized | Control | Raw | t-test p-value | Flags
1 Patient ID 112 , CHF Etiology Nonfailing 2.687 193.2 519.3 no replicates P
2 Patient ID 118 , CHF Etiology Nonfailing 2.968 172.8 512.7 no replicates P
3 Patient ID 148 , CHF Etiology Nonfailing 1.973 220.9 435.8 no replicates P
4 Patient ID 200 , CHF Etiology Nonfailing 2.92 177.3 517.7 no replicates P
5 Patient ID 249 , CHF Etiology Nonfailing 1.86 180.5 335.6 no replicates P
6 Patient ID 291 , CHF Etiology Nonfailing 2.884 196 567.2 no replicates P
7 Patient ID 294 , CHF Etiology Nonfailing 2.931 168.2 493 no replicates P
8 Patient ID 300 , CHF Etiology Nonfailing 1.998 196.5 392.7 no replicates P
9 Patient ID 322 , CHF Etiology Nonfailing 2.511 183.7 461.4 no replicates P
10 Patient ID 325 , CHF Etiology Nonfailing 2.479 186.3 461.8 no replicates P
11 Patient ID 326 , CHF Etiology Nonfailing 4.493 189.8 852.6 no replicates P
Normalized Intensity (log scale)
100
10
1
0.1
0.01
Patient ID
112 148 249 294 322 326 115 233 267 297 342 464 102 206 225 317 348 64 82 85 93 98
Nonfailing
Ischemic
Idiopathic
Tools
Minimum correlation 0.95
Find Similar
Complex Correlations
Save Expression Profile
Lists Containing 227884_at
Gene Ontology (GO SLIMS)
GO SLIMS Biological Process
cell communication
cell-cell signaling (GO:0007267)
cell growth and-or maintenance
Web Links
GenBank (GB): AW296067
Ensembl : TAF15
LocusLink: AW296067
OMIM: TAF15
Netaffx: 227884_at
UniGene (UG): Hs.402752
Search...
Close
Help
(4 of 6)
<<
>>

Figure 7-3 Gene Inspector Window

In the center of the window is a browser showing a graph of normalized values for the gene across all conditions. At the bottom of the window, from left to right, are gene inspection tools, lists containing your genes, and web links to databases.

Gene Identification Section

Annotations for the selected gene from the Master Table of Genes are displayed in the upper left corner of the Gene Inspector window in the Gene Identification section.

Other Notes Section

The Other Notes section contains a field in which you can add notes about the genes. To save the notes, click the **Save Notes** button.

Data Table Section

The table in the upper-right corner of the Gene Inspector is the Data Table. It contains the following information:

- **Condition—**Shows the condition under which the measurement was taken.
- **Normalized—**Shows the normalized data value. For details on normalization, go to Chapter 10, "Normalizing Data".
- **Control—**Shows the control strength for the gene. For more information about control strengths, go to "Per-chip Normalizations" on page 10-15.
- **Raw—**Shows the raw value of the data, just as it came off the chip or out of the scanner, averaged over the replicates (if any) in the given condition.
- **t-test p-value—**Shows the measure of the likelihood of this gene's expression value being different from one, assuming the data is centered around one. The t-test p-value is applicable only to replicated data.
- **Flags—**Indicates if your data are reliable. Whether or not you have flags depends on your instrumentation and what you have entered into your master gene table. Go to "Measurement Flags" on page 10-27 for more information.

T-test P-value

In cases where there is replicate data, a one-sample Student's t-test is calculated to test whether the mean normalized expression value for the gene is statistically different from 1.0. The t-statistic is calculated as:

$$t = \frac{\bar{X} - 1}{S_x / (\sqrt{n})}$$

where $\bar{X} = \sum_{i-1}^{n} X_i$ is the sample average of the n normalized expression levels $X_1,\ldots,X_n$,

and $S_x = \sqrt{\frac{1}{n-1}} \sum_{i-1}^{n} (X_i - \bar{X})^2$

is the sample standard deviation of the replicates. The value of t is compared with a table of the Student's t-distribution with $n - 1$ degrees of freedom to yield the significance level (or p-value) for a two-sided test that the mean gene intensity differs significantly from 1.0.

Browser Display

The Browser displays shows you a graph of the gene's expression values for the current experiment interpretation. Right-click on the browser to open the Display Options window and specify the display options you want.

For details on the other options, see "Using the Genome Browser" on page 7-1.For information about error bars, see "Using Cross-Gene Error Models" on page 5-73. For information about creating a resizable picture, see "Printing Images" on page 17-4.

Gene Inspection Tools

The box in the bottom left corner of the Gene Inspector window contains tools allowing you to search for genes having similar expression profiles to the gene currently displayed.

- **Find Similar—**lets you search for genes with similar expression profiles to the gene being inspected. Each gene expression profile must have the required minimum correlation to be considered similar. The higher the minimum correlation (maximum 1), the closer the gene expression profiles must be. Enter this number in the Minimum correlation box above the **Find Similar** button. For information on using the Find Similar function, see "Inspecting Gene Lists" on page 12-11.
- **Complex Correlation—**lets you make a gene list comparing the gene being inspected to genes having similar expression profiles in multiple experiments, with more complex parameters than the Find Similar tool allows. For information on using the Complex Correlation function, see "Correlation of Expression Change" on page 14-11.
- **Save As Expression Profile—**lets you save your gene's expression values as an expression profile, which you can use to make lists. For information on making lists from expression profiles, see "Making Gene Lists from Expression Profiles" on page 12-21.

Lists Containing Your Gene Section

In the bottom center of the Gene Inspector window is a Navigator for the lists containing your gene. Double-click on a list to open the Gene List Inspector window. For information about this window, see "Using the Gene List Inspector" on page 7-33.

Searching Internet Databases

In the Windows version of GeneSpring, you can set up the Gene Inspector window to search public databases. In the Web Links panel (Figure 7-4) of the Gene Inspector window, select a public database and click the **Search** button.

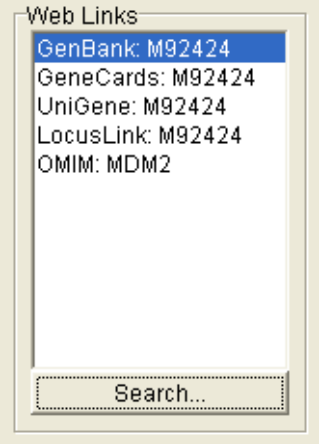


Figure 7-4 Gene Inspector: Web Links Panel

To configure a web browser with a Macintosh, go to **Edit > Preferences > Browser** and enter the appropriate pathway.

Using the Sample Inspector

The Sample Inspector lets you view and edit detailed data about the selected sample.

Opening the Sample Inspector

To open the Sample Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Experiments** folder, open the **All Samples** folder, right-click a sample, and then select the **Inspect** command.
  - In the **Navigator**, open the **Experiments** folder, open the **All Samples** folder, and then double-click a sample.
  - Select **Experiments > Sample Manager**, right-select a sample, and then select the **Inspect** command.
  - Select **Experiments > Sample Manager**, and then double-click a sample.

Sample Inspector Window

The Sample Inspector window (Figure 7-5) has two sections.The upper section contains basic information about a sample.

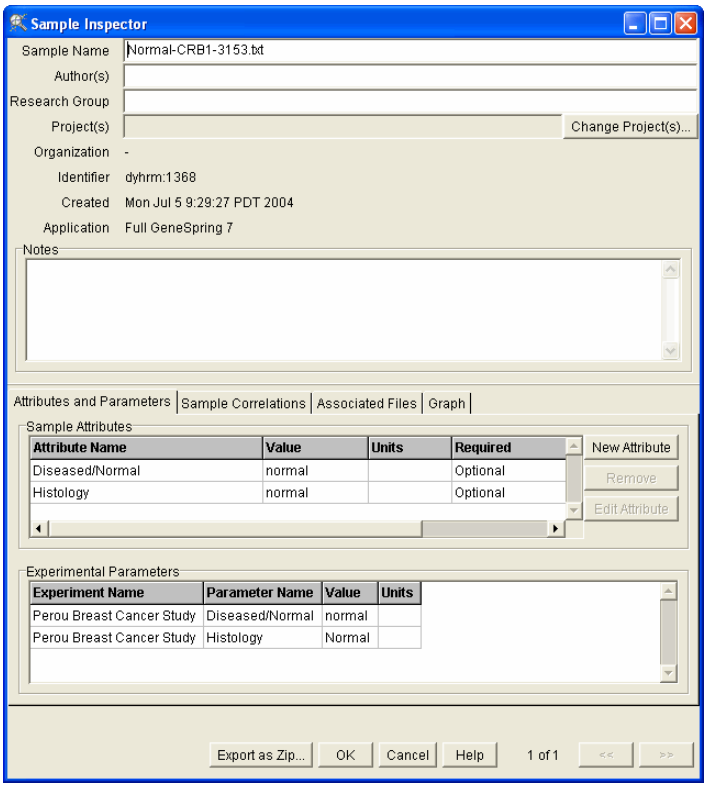


Figure 7-5 Sample Inspector Window

Summary Information Section

- **Name**—Show the name of the sample.
- **Author(s)**—Lets you enter the name of the authors associated with the sample.
- **Research Group**—Lets you enter the name of the research group associated with the sample.
- **Project**—Shows the name of the project assigned to the sample.
- **Organization**—Shows the name of the organization associated with the sample. This information cannot be edited; it is derived from the license key.
- **Identifier**—Shows the sample identifier.
- **Created**—Shows the date and time at which the sample was created in GeneSpring.
- **Application**—Shows the application used to create the sample.
- **Notes**—Lets you enter information about the sample.

Buttons

The Sample Inspector window contains the following buttons:

- **Change Projects**—Assigns the selected sample to a project. Go to "Working with Projects" on page 7-40 for more information.
- **Export as Zip—**Saves the sample as a compressed file.
- **OK**—Saves your changes and exits the window.
- **Cancel**—Closes the Sample Inspector without saving any changes.
- **Help**—Displays online help.
- << >>—Navigates backward or forward through samples.

Attributes and Parameters Tab

In the Attributes and Parameters tab (Figure 7-6), two tables appear: Sample Attributes and Experimental Parameters.

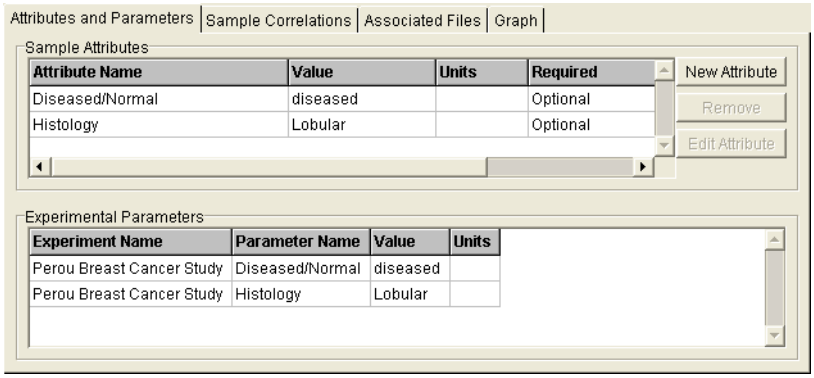


Figure 7-6 Sample Inspector Window: Attributes and Parameters Tab

In the Sample Attributes table, you can view the attributes of the sample being inspected. You can also add, remove, or edit any of these attributes.

The Experimental Parameters table displays parameters assigned to this sample in all relevant experiments. Since a sample may be part of several experiments, you cannot edit experiment parameters from this window. For information on how to edit experiment parameters, see "Setting Up Experiment Parameters" on page 5-37.

Sample Attribute Window

The Sample Attribute window (Figure 7-7), which is opened by clicking the New Attribute button, lets you add or edit a new sample attribute.

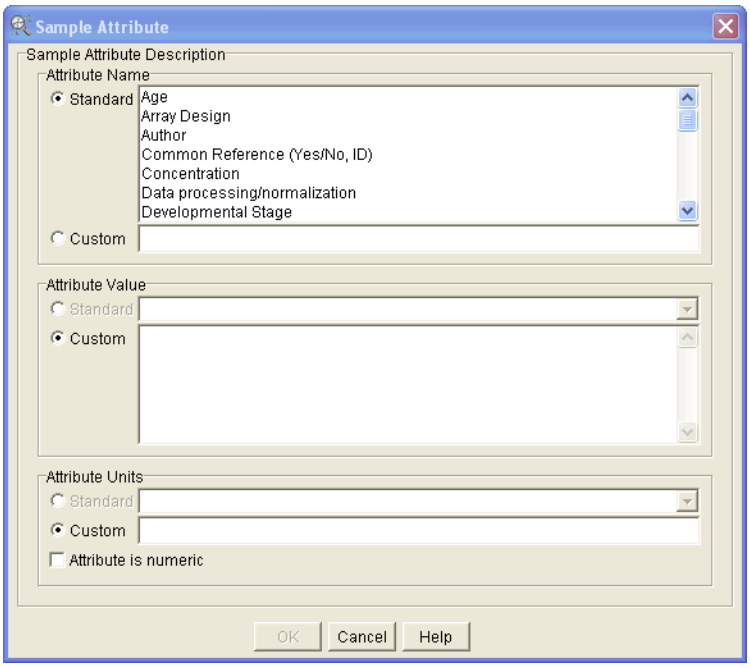


Figure 7-7 Sample Attribute Window

You can specify the following information in this window:

- **Attribute Name**—Shows the name of the attribute being added. You can select a standard attribute name from the scrolling list, or select the **Custom** option and enter a new attribute name.
- **Attribute Value**—Shows the value of the attribute being added. For many standard attributes, there is a default value. You can accept the default by choosing the **Standard** option, or you can select the **Custom** option and enter a new value.
- **Attribute Units**—Shows the units in which the attribute is measured. For many standard attributes, there are standard units. You can accept one of the standard units from the scrolling list, or you can select the **Custom** option and enter new units. If the units are numeric, select the **Attribute is numeric** box.

When you are done, click **OK** to save your new attribute and return to the Sample Inspector.

To remove an attribute, select it in the **Sample Attributes** list and click **Remove**.

To edit an attribute, select it in the **Sample Attributes** list and click **Edit Attribute**. The Sample Attribute window opens. Proceed as you would if you were adding a new attribute.

Sample Correlations Tab

The Sample Correlations tab (Figure 7-8) shows the correlation between the current sample and the samples in the currently selected experiment. This list appears only when an experiment that contains the sample being inspected is selected.

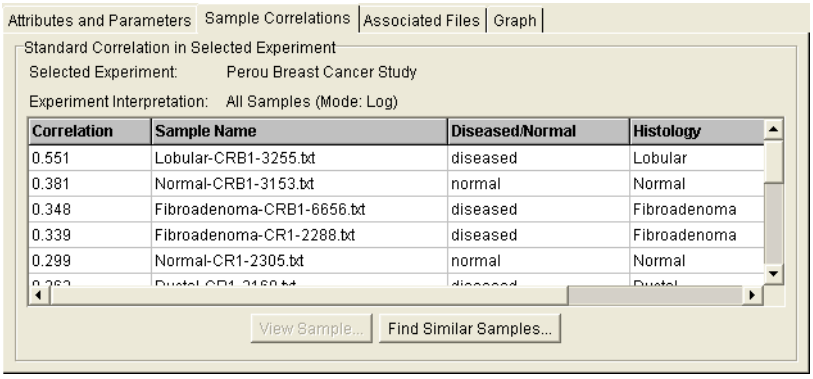


Figure 7-8 Sample Inspector Window: Sample Correlations Tab

You can click any of the samples in this list and view them in another Sample Inspector window by clicking **View Sample**.

Associated Files Tab

The Associated Files tab (Figure 7-9) lists any files that may be associated with the sample, including data files, array images, sample images, etc. If there is a sample image included among the associated files, it is displayed in the **Sample Image** panel to the right of the list.

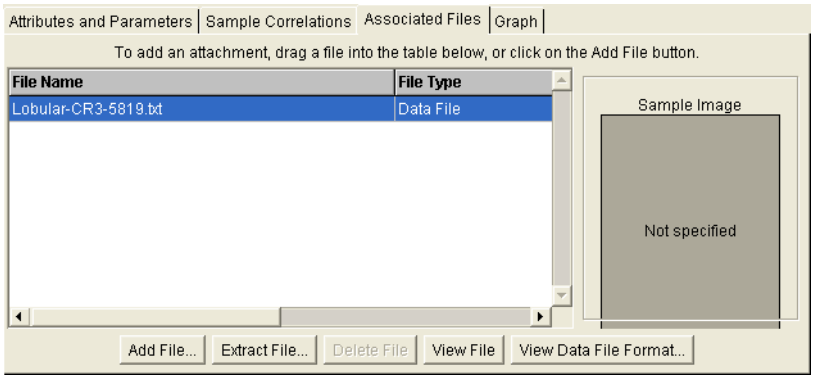


Figure 7-9 Sample Inspector Window: Associated Files Tab

You can re-order the files in this list by clicking the property to sort by in the list headers. For example, to sort by file type rather than file name, click the **File Type** column header.

From this window you perform the following functions:

- **Add File**—To add a file, click **Add File**, select the file you want from the Select a file to attach dialog, and then click the **Open** button. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the file you want from the list and click **Extract File**. Choose a location from the Extract File dialog and click the **Save** button. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. The appropriate program to display is selected automatically if the file type is known.
- **View Data File Format**—Click to display the column assignments for the selected data file as it was loaded into GeneSpring.

Graph Tab

The Graph view (Figure 7-9) is available only if an experiment containing the current sample is selected. It displays a graph of the raw sample data.

Note: The Graph tab is not available on MacOS.

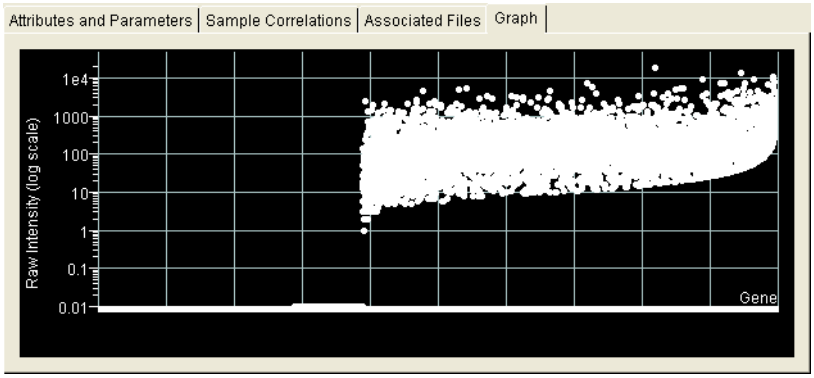


Figure 7-10 Sample Inspector Window: Graph Tab

Using the Experiment Inspector

The Experiment Inspector lets you inspect the details associated with experiments.

Opening the Experiment Inspector

To open the Experiment Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Experiments** folder, right-click an experiment, and then select the **Inspect** command.
  - In the **Navigator**, open the **Experiments** folder, and then double-click an experiment.

Experiment Inspector Window

The lower section of the Experiment Inspector window (Figure 7-11) provides information about the selected experiment.

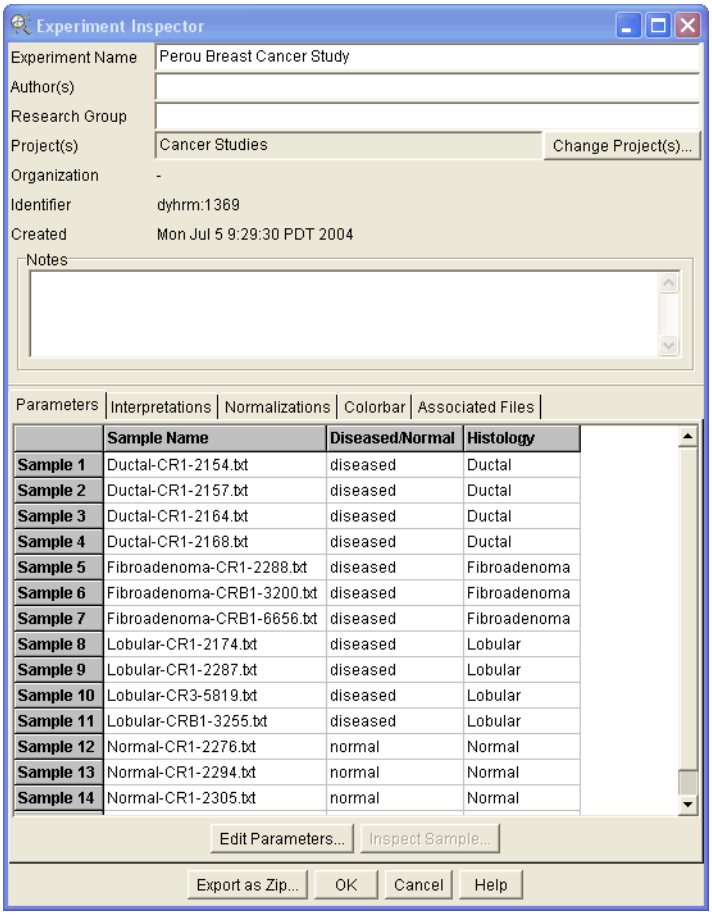


Figure 7-11 Experiment Inspector Window

Summary Information Section

- **Name**—Show the name of the experiment.
- **Author(s)**—Lets you enter the name of the authors associated with the experiment.
- **Research Group**—Lets you enter the name of the research group associated with the experiment.
- **Project**—Shows the name of the project assigned to the experiment.
- **Organization**—Shows the name of the organization associated with the experiment. This information cannot be edited; it is derived from the license key.
- **Identifier**—Shows the experiment identifier.
- **Created**—Shows the date and time at which the experiment was created in GeneSpring.
- **Notes**—Lets you enter information about the experiment.

Buttons

The Experiment Inspector window contains the following buttons:

- **Change Projects**—Assigns the selected experiment to a project. Go to "Working with Projects" on page 7-40 for more information.
- **Export as Zip—**Saves the experiment as a compressed file.
- **OK**—Saves your changes and exits the window.
- **Cancel**—Closes the Experiment Inspector without saving any changes.
- **Help**—Displays online help.

Parameters Tab

On the Parameters tab (Figure 7-12), you can view the experiment parameters and their possible values.

Parameters | Interpretations | Normalizations | Colorbar | Associated Files

	Sample Name	Diseased/Normal	Histology
Sample 1	Ductal-CR1-2154.txt	diseased	Ductal
Sample 2	Ductal-CR1-2157.txt	diseased	Ductal
Sample 3	Ductal-CR1-2164.txt	diseased	Ductal
Sample 4	Ductal-CR1-2168.txt	diseased	Ductal
Sample 5	Fibroadenoma-CR1-2288.txt	diseased	Fibroadenoma
Sample 6	Fibroadenoma-CRB1-3200.txt	diseased	Fibroadenoma
Sample 7	Fibroadenoma-CRB1-6656.txt	diseased	Fibroadenoma
Sample 8	Lobular-CR1-2174.txt	diseased	Lobular
Sample 9	Lobular-CR1-2287.txt	diseased	Lobular
Sample 10	Lobular-CR3-5819.txt	diseased	Lobular
Sample 11	Lobular-CRB1-3255.txt	diseased	Lobular
Sample 12	Normal-CR1-2276.txt	normal	Normal
Sample 13	Normal-CR1-2294.txt	normal	Normal
Sample 14	Normal-CR1-2305.txt	normal	Normal

Edit Parameters... | Inspect Sample...

Figure 7-12 Experiment Inspector Window: Parameters Tab

Click **Edit Parameters** to view the Experiment Parameters window (Figure 7-13), which allows you to modify experiment parameters.

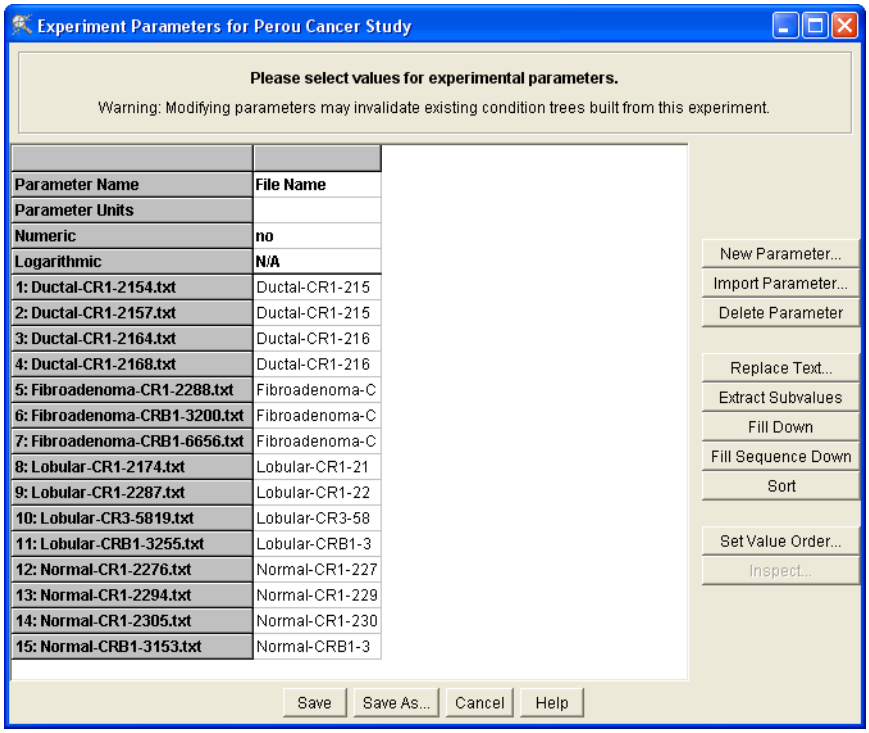


Figure 7-13 Experiment Parameters Window Tab

See "Setting Up Experiment Parameters" on page 5-37 for details on this window. When you click **OK**, any changes you make are saved and applied to your experiment.

To view details on a particular sample, select it in the list and click **Inspect Sample**. Alternatively, you can double-click a sample. This opens the Sample Inspector window. For more information about the Sample Inspector, see "Using the Sample Inspector" on page 7-19.

Interpretations Tab

The Interpretations tab (Figure 7-14) lets you view all the interpretations associated with the selected experiment. Click on an interpretation in the list to select it. To edit an interpretation, double-click it or select it in the list and click **Edit Interpretation**. The Change Interpretation window opens.

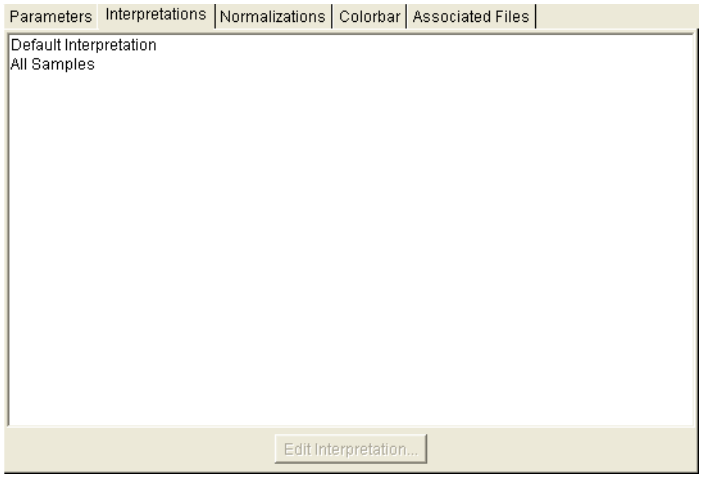


Figure 7-14 Experiment Inspector Window: Interpretations Tab

Normalizations Tab

The Normalizations tab lets you view the normalizations currently being used in your experiment. To edit normalizations, click **Edit Normalizations**. The Experiment Normalizations window opens. See "Experiment Normalizations Window" on page 10-2 for details on this window. Click **OK** to save your changes.

To view a detailed description of all normalization applied in this experiment, click the **View Text Description** button. A window opens with a description of the selected normalization. You can copy the text in this window by clicking Copy to Clipboard. You can then paste the text into a text editor.

Colorbar Tab

The Colorbar tab (Figure 7-15) lets you view and edit the default range of expression in your experiment's coloration scheme. You can also specify whether or not to show trust on the Colorbar by clicking the option next to your preferred choice.

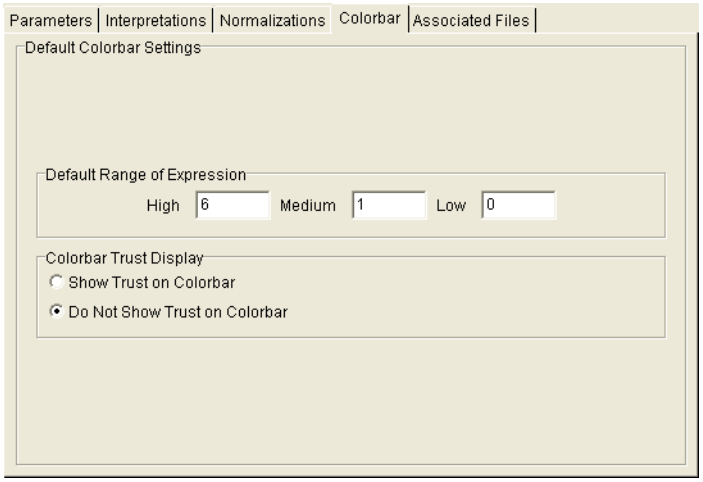


Figure 7-15 Experiment Inspector Window: Colorbar Tab

Use this tab to view the default range of expression in your experiment's coloration scheme. You can also see whether or not trust is shown on the Colorbar by clicking the option next to your preferred choice.

For more information coloration in your experiment, see "Changing the Colorbar Range" on page 8-63.

Associated Files Tab

The Associated Files tab (Figure 7-16) shows files that are associated with the experiment, like published papers, documentation, etc.

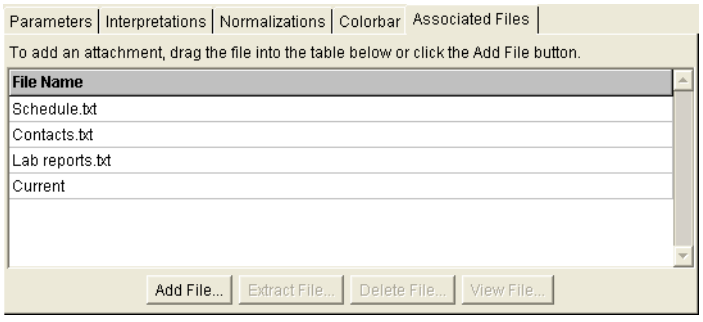


Figure 7-16 Experiment Inspector Window: Associated Files Tab

From this tab you can perform the following procedures:

- **Add File**—To add a file, click **Add File**, select the file you want from the **Select a file to attach** dialog, and then click the **Open** button. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the file that you want from the list and click **Extract File**. Choose a location from the **Extract File**

dialog and click the **Save** button. This does not remove the file from your list. It simply places a copy of the file in a new location.

- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. GeneSpring automatically selects the appropriate program to display the file if the file type is known.

Using the Condition Inspector

A condition is a unique combination of parameters as applied to your sample. Each condition may be a single sample or a group of replicate samples combined based upon the parameter values defined for each sample. The easiest way to think of this is as the parameters under which the sample(s) was observed. If you have no replicates, condition and sample can be considered synonymous.

Opening the Condition Inspector

To open the Condition Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Experiment** folder, open the **Interpretation** folder, right-click a condition, and then select the **Inspect** command.
  - In the **Navigator**, open the **Experiment** folder, open the **Interpretation** folder, and then double-click a condition.

Condition Inspector Window

The Condition Inspector window (Figure 7-17) provides detailed information about the selected experiment condition.

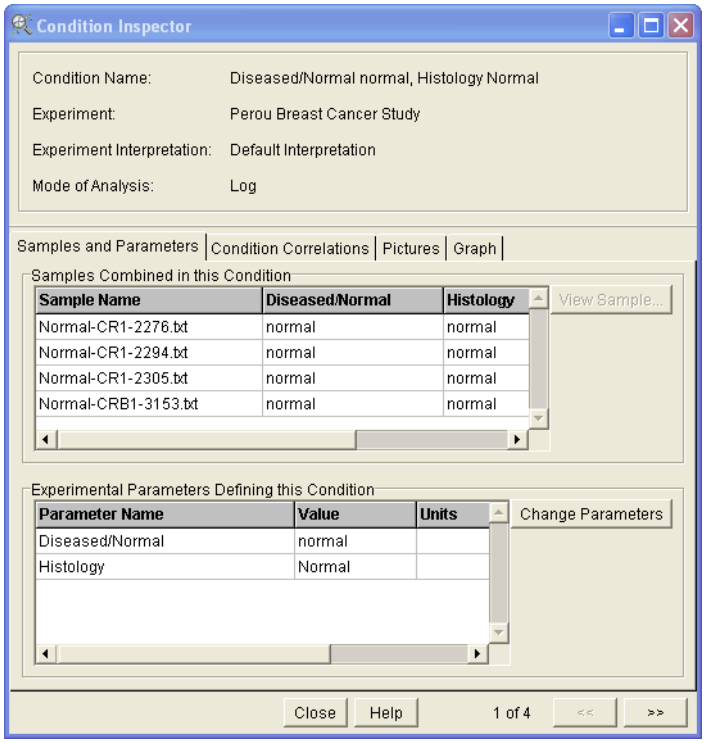


Figure 7-17 Condition Inspector Window

Summary Information Section

- **Condition Name**—Shows the name of the condition.
- **Experiment**—Shows the experiment associated with the condition.
- **Experiment Interpretation**—Shows the experiment interpretation associated with the condition.
- **Mode of Analysis**—Shows the mode of analysis associated with the experiment.

Buttons

The Condition Inspector window contains the following buttons:

- **Close**—Closes the Condition Inspector without saving any changes.
- **Help**—Displays online help.
- << >>—Navigates backward or forward through samples.

Samples and Parameters Tab

The Sample and Parameters tab (Figure 7-18) contains two sections:

- **Samples Combined in this Condition**—Lists the samples in the selected condition. Select a sample and click **View Sample** to invoke the Sample Inspector for that sample. See "Using the Sample Inspector" on page 7-19 for more information.

- **Experimental Parameters Defining this Condition—**Lists the parameters associated with this condition. To edit parameters, click Change Parameters. For more information on the Change Parameters window, see "Setting Up Experiment Parameters" on page 5-37.

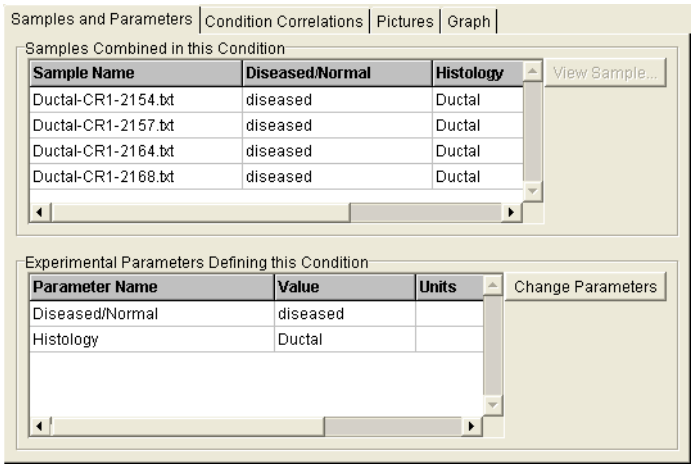


Figure 7-18 Condition Inspector Window: Samples and Parameters Tab

Condition Correlations Tab

The Condition Correlations tab (Figure 7-19) contains a list of all other conditions in the experiment, along with columns corresponding to their associated parameters.

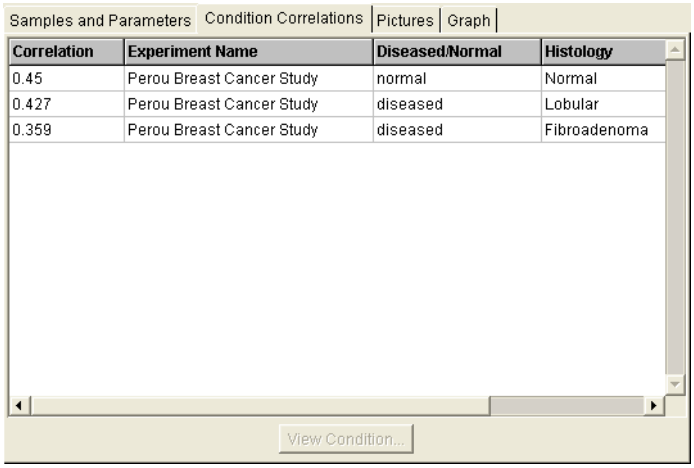


Figure 7-19 Condition Inspector Window: Condition Correlations Tab

The Correlation column shows how closely correlated the other conditions in the experiment are to the one under inspection. The conditions are listed from most closely correlated to least correlated. This feature uses "standard correlation" to measure the similarity of the selected condition and all others in the experiment. This cannot be changed. To use another metric, you must create a script using the "Condition Correlation" building block. For details on creating scripts, see "Working with Scripts" on page 16-1.

Pictures Tab

The Pictures tab displays the sample images, if any, associated with this condition. Double-click an image to view it at its full size.

Graph Tab

The Graph tab (Figure 7-20) displays the condition in graph form.

Note: The Graph tab is not available on MacOS.

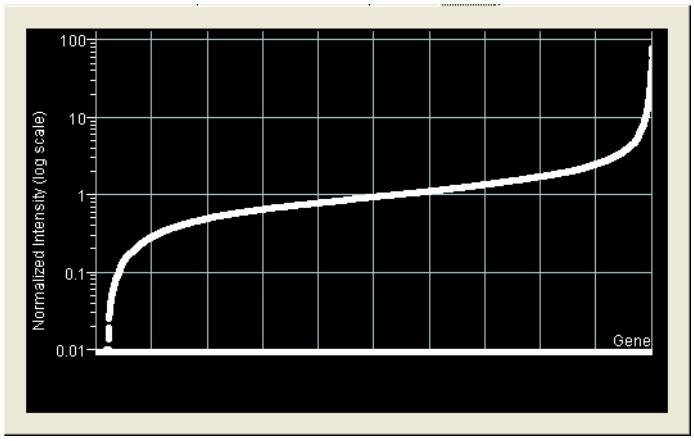


Figure 7-20 Condition Inspector Window: Graph Tab

Using the Gene List Inspector

You can view the contents of a gene list and its creation method using the Gene List Inspector window. This window is especially useful for learning about lists identified using the Similar List function.

Opening the Gene List Inspector

To open the Gene List Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Inspect** command.
  - In the **Navigator**, open the **Gene Lists** folder, and then double-click the gene list you want.

Gene List Inspector Window

The Gene List Inspector window (Figure 7-21) displays descriptive and illustrative information about the selected gene list.

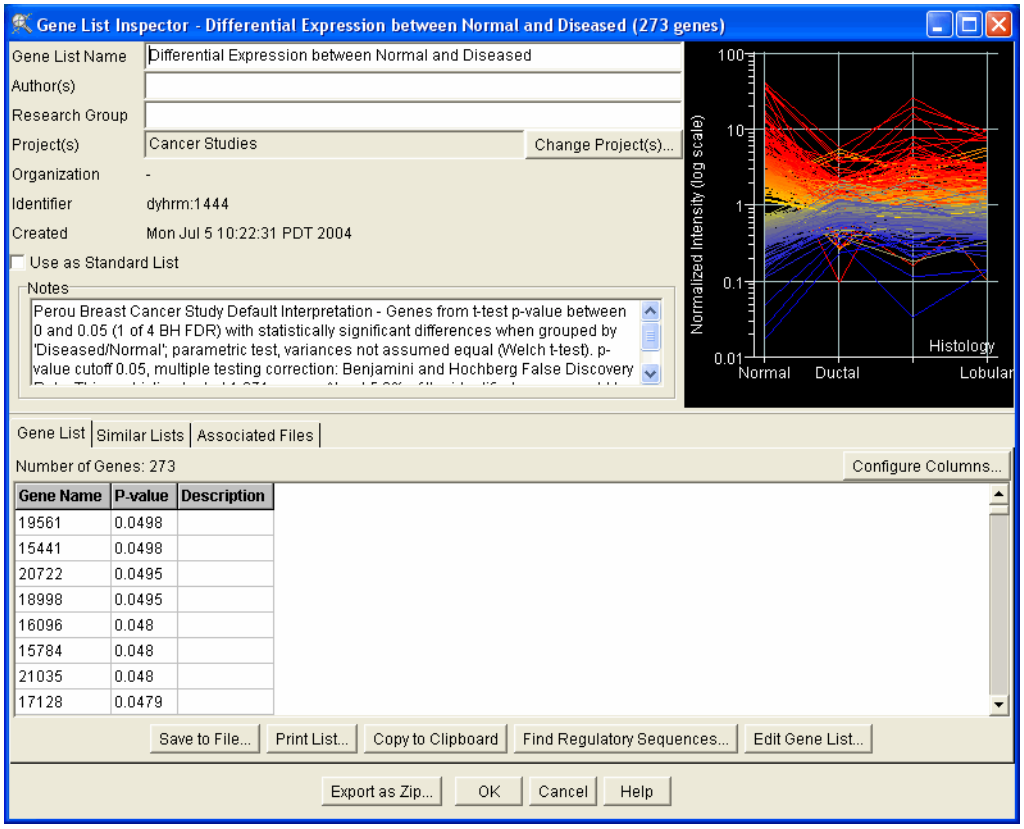


Figure 7-21 Gene List Inspector Window

Summary Information Section

- **Name**—Show the name of the gene list.
- **Author(s)**—Lets you enter the name of the authors associated with the gene list.
- **Research Group**—Lets you enter the name of the research group associated with the gene list.
- **Project**—Shows the name of the project assigned to the gene list.
- **Organization**—Shows the name of the organization associated with the gene list. This information cannot be edited; it is derived from the license key.
- **Identifier**—Shows the gene list identifier.
- **Created**—Shows the date and time at which the gene list was imported into GeneSpring.
- **Notes**—Lets you enter information about the gene list.

Use as Standard List Option

A standard list is a special type of gene list in GeneSpring. GeneSpring automatically identifies some gene lists as standard lists. For example, the GO SLIMS ontology gene list represents a special type of gene list that GeneSpring identifies as a standard list.

You can also manually save a gene list as a standard list by selecting the Use as Standard List option. Once you save a gene list as a standard list, you can limit searches in the Gene List Inspector window to only standard lists, all lists, or no lists (which effectively disables the Similar List function). Go "Miscellaneous Tab" on page 3-24 for configuration information.

Genome Browser

In the upper right corner of the window is a browser graphing your list. Right-click on the graph for a menu of options. See "Using the Genome Browser" on page 7-1 for information on browser options.

Buttons

The Gene List Inspector window contains the following buttons:

- **Change Projects**—Assigns the selected gene list to a project. Go to "Working with Projects" on page 7-40 for more information.
- **Export as Zip—**Saves the gene list as a compressed file.
- **OK**—Saves your changes and exits the window.
- **Cancel**—Closes the Gene List Inspector without saving any changes.
- **Help**—Displays online help.

Gene Lists Tab

The Gene List tab (Figure 7-22) displays a table of all the genes included in the selected list. Double-click a gene or cell in this table to view a Gene Inspector window for the selected gene. See "Using the Gene Inspector" on page 7-16 for information on the Gene Inspector window. Click on any column header in the displayed gene list to sort the table by the values in that column.

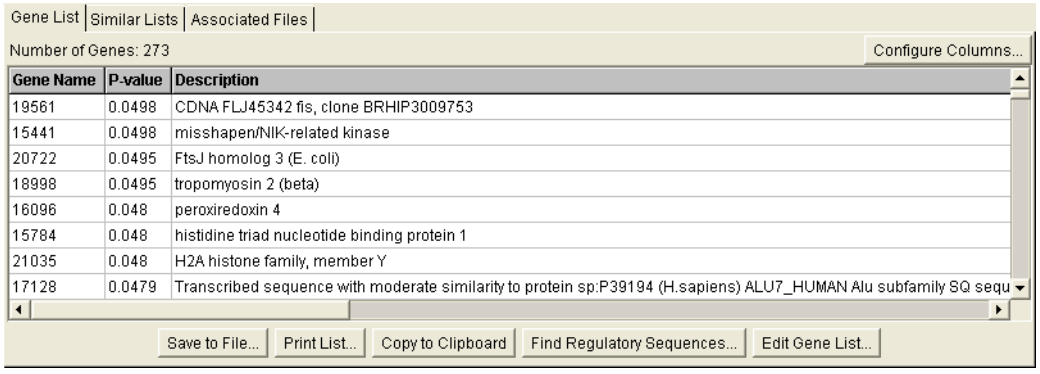

Gene List
Similar Lists
Associated Files
Number of Genes: 273
Configure Columns...
Gene Name
P-value
Description
19561 0.0498 CDNA FLJ45342 fis, clone BRHIP3009753
15441 0.0498 misshapen/NIK-related kinase
20722 0.0495 FtsJ homolog 3 (E. coli)
18998 0.0495 tropomyosin 2 (beta)
16096 0.048 peroxiredoxin 4
15784 0.048 histidine triad nucleotide binding protein 1
21035 0.048 H2A histone family, member Y
17128 0.0479 Transcribed sequence with moderate similarity to protein sp:P39194 (H.sapiens) ALU7_HUMAN Alu subfamily SQ sequ
Save to File...
Print List...
Copy to Clipboard
Find Regulatory Sequences...
Edit Gene List...

Figure 7-22 Gene List Inspector Window: Gene List Tab

From this tab, you have the following options:

- **Configure Columns**—Selects which columns to display in the Gene Lists tab. You can choose from any of the columns in your Master Table of Genes except the Sequence column.
- **Save to File**—Saves the entire gene list as a tab-delimited file.
- **Print List**——Sends the selected gene list to a printer for printing.
- **Copy to Clipboard**—Copies the contents of the gene list to the clipboard. You can then paste the list into another application, such as a text editor or spreadsheet.
- **Find Regulatory Sequences**—Opens the Find Potential Regulatory Sequences window with the current gene list pre-selected. This button is available only if the genome is fully sequenced. For more information on this window, see "Finding Potential Regulatory Sequences" on page 14-1.
- **Edit Gene List**—Opens the Gene List Editor window. For more information, see "Creating or Editing Gene Lists" on page 12-5.

Similar Lists Tab

The Similar Lists tab (Figure 7-23) displays names of lists resembling the selected list, or containing a statistically significant number of overlapping genes. The overlap is calculated using a standard Fisher's exact test and the p-value is adjusted with a Bonferroni multiple testing correction.

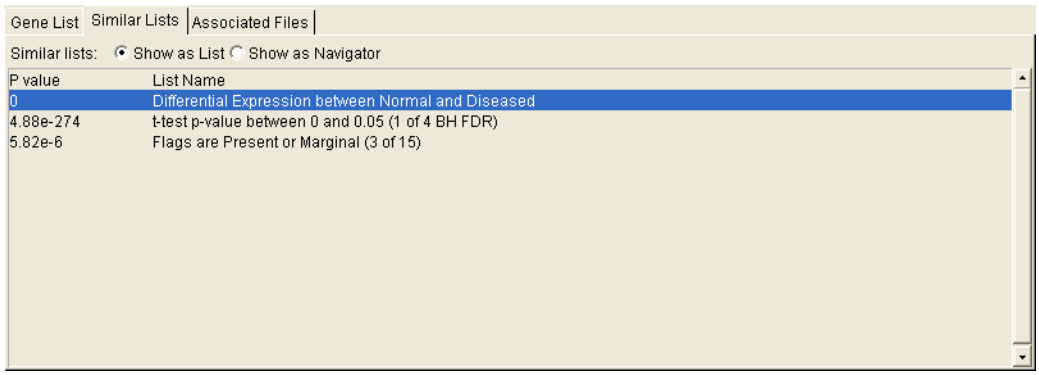


Figure 7-23 Gene List Inspector Window: Similar Lists Tab

Two methods are available to view these lists:

- **List View**—Displays a simple two-column list. In this view, statistical significance is listed as the p-value for each of the similar lists.
- **Navigator View**—Displays a Navigator-style listing.

Right-click a list to print or copy. Double-click a list to view a Gene List Inspector window for that list.

Associated Files Tab

The Associated Files tab (Figure 7-24) lists any files that are associated with the selected gene list, such as research publications, documentation, etc.

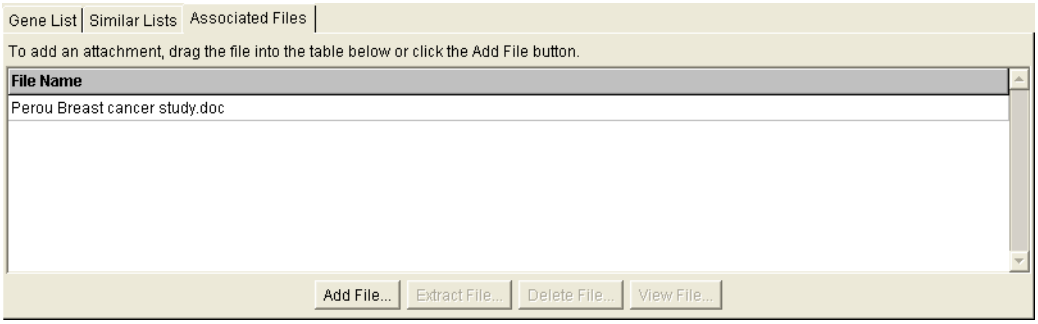


Figure 7-24 Gene List Inspector Window: Associated Files Tab

From this tab, you can perform the following tasks:

- **Add File**—To add a file, click **Add File**, select the file you want from the **Select a file to attach** dialog, and then click the **Open** button. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the file you want from the list and click **Extract File**. Choose a location from the **Extract File** dialog and click the **Save** button. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. The appropriate program is automatically selected if the file type is known.

Using the Classification Inspector

The Classification Inspector lets you learn about the method used to construct a classification, or the variability explained by each class within a classification.

Opening the Classification Inspector

To open the Classification Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Classifications** folder, right-click a classification, and then select the **Inspect** command.
  - In the **Navigator**, open the **Classifications** folder, and then double-click a classification.

Classification Inspector Window

The Classification Inspector window (Figure 7-25) contains information about the method used to make the classification. The top part of the window contains descriptive information, such as the classification name, authors, organization, creation date, etc.

If the classification is the result of clustering, the Notes field displays information such as the type of clustering, the distance metric, and the number of iterations that were used to perform the clustering. You can also add your own comments about the classification here for future reference.

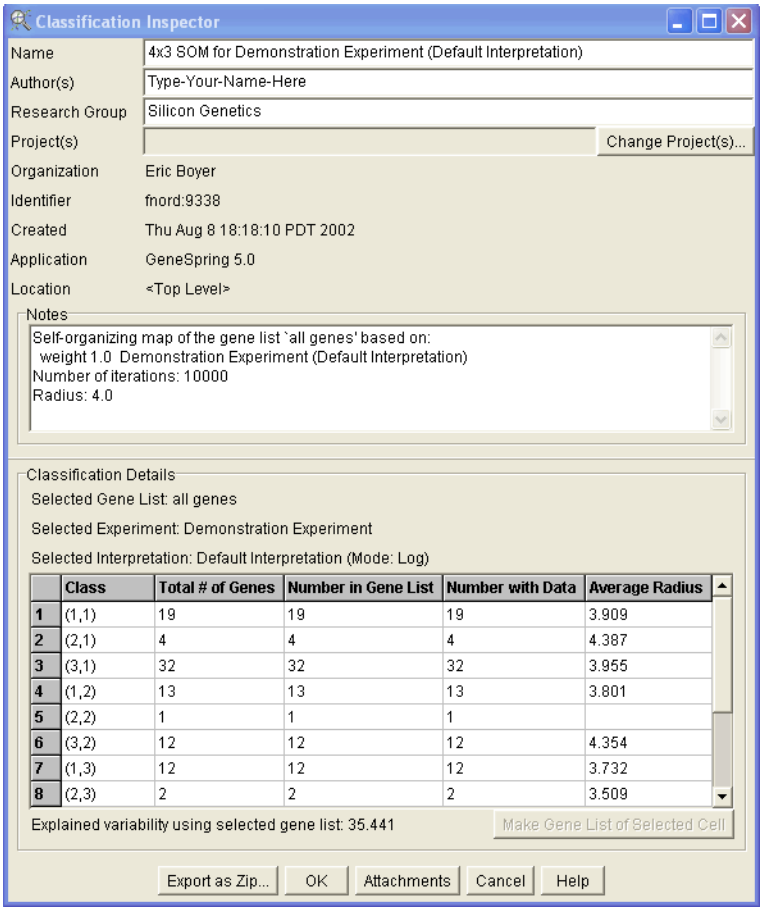


Figure 7-25 Classification Inspector

Summary Information Section

- **Name**—Show the name of the classification.
- **Author(s)**—Lets you enter the name of the authors associated with the classification.
- **Research Group**—Lets you enter the name of the research group associated with the classification.
- **Project**—Shows the name of the project assigned to the classification.

- **Organization**—Shows the name of the organization associated with the classification. This information cannot be edited; it is derived from the license key.
- **Identifier**—Shows the classification identifier.
- **Created**—Shows the date and time at which the classification was created in GeneSpring.
- **Application**—Shows the application used to create the classification.
- **Location**—Shows the name of the directory in which the classification is saved.
- **Notes**—Lets you enter information about the classification.

Buttons

The Classification Inspector window contains the following buttons:

- **Change Projects**—Assigns the selected classification to a project. Go to "Working with Projects" on page 7-40 for more information.
- **Export as Zip—**Saves the classification as a compressed file.
- **OK**—Saves your changes and exits the window.
- **Attachments—**Attaches a file or folder to the classification.
- **Cancel**—Closes the Classification Inspector without saving any changes.
- **Help**—Displays online help.

Classification Details Table

The bottom half of the Classification Inspector contains a table with the following columns:

- **Class—**Shows the name given to each class.
- **Number in Gene List—**Shows the number of genes in each class, which are present in the currently selected gene list.
- **Number with Data—**Shows how many of the genes in the previous column have expression data in the currently selected interpretation.
- **Average Radius—**Shows the root mean square of the Euclidean distances between each gene and the centroid of each class. Classes with large radii are spread out and classes with small radii are tightly grouped.

Any of the collections of genes listed in the table can be examined by selecting a table cell with a number of genes, and clicking the **Make Gene List of Selected Cell** button.

Percent Explained Variability

The percent value of the explained variability for the genes in the current gene list is shown below the table. The explained variability formula weights classes by the number of genes in each class. The variables in the formula are:

- *c*—The number of classes (including "unclassified", but not "no data").
- *n*—The total number of genes with data.

- n_I—The number of genes with data in class *i.*
- D_i —The distance of each class centroid from the overall data centroid.
- d_{ij}—The distance of each gene from the centroid of its class.

The formula is:

$$B = \sum_{i-1}^{c} n_i D_i^2$$

$$W = \sum_{i-1}^{c} \sum_{j-1}^{n_i} d_{ij}^2$$

$$E = 100 \times max\left(1 - \frac{W/(n-c)}{(B+W)/(n-1)}, 0\right) \quad (\text{If } c \geq n \text{ then } E=0$$

E is the percent variability explained.

Note that the percent explained variability depends on the selected experiment, and the selected gene list. It is calculated using Euclidean distance of the gene expression profiles of the conditions interpreted in the interpretation made (ratio, log of ratio, or fold change).

References for the Classification Inspector

Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics, 3, 1-27.

Gordon, A. D. *Classification*, 2nd Ed. Monographs on Statistics and Applied Probability 82. Chapman & Hall/CRC, Boca Raton (1999).

Working with Projects

This section explains how to use projects to manage data objects in GeneSpring. It covers the following topics:

- Project Functions
- Project Assignment
- Filter Navigator Window
- Assigning Data Objects to Projects
- Assigning New Experiments to Projects

Project Functions

To provide greater control over navigating, filtering, and managing data objects in the GeneSpring window, GeneSpring 7 introduces the concept of a project. A project is a virtual container for data objects that represents the way data is structured in laboratory notebooks, internal reports, or proprietary databases.

Projects enable you to tag data objects associated with an experiment using your own labels so they can be identified, searched, and filtered using the GeneSpring search and filter functions. For example, projects let you:

- Determine what experiments were conducted to address the objectives of the study
- Determine what samples were used in the creation of the experiments
- Identify data objects derived from the analysis of individual experiments
- Navigate to data objects in GeneSpring and Signet

A project can include multiple experiments and it can cover multiple genomes. Separate experiments can also share data objects used in common.

Project Assignment

You can assign the following data objects to projects:

- Gene Lists
- Experiments
- Gene Trees
- Condition Trees
- Classifications
- Pathways
- Array Layouts
- Expression Profiles
- Bookmarks
- Samples

Filter Navigator Window

After the data objects have been tagged, you can use the Filter Navigator window (Figure 7-26) to locate data objects that have been assigned to projects, or those that have not yet been assigned to projects.

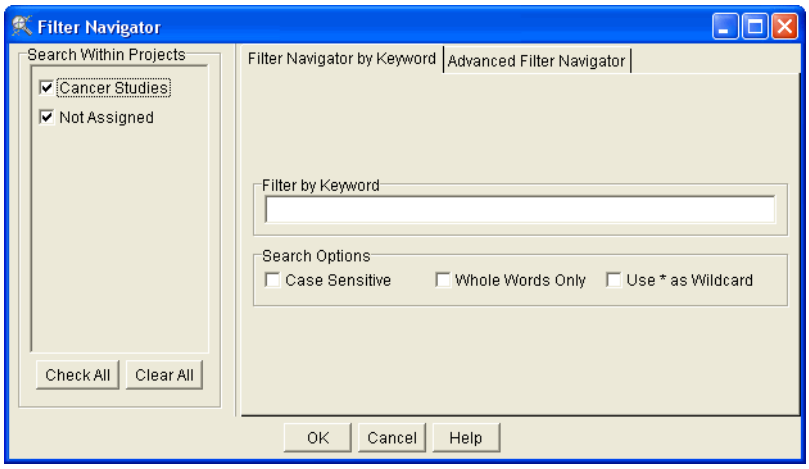


Figure 7-26 Filter Navigator Window

The following section describes the elements in the Filter Navigator.

Filter Navigator By Keyword Tab

The Filter Navigator by Keyword tab (Figure 7-27) is the default tab that appears when the Filter Navigator is opened.

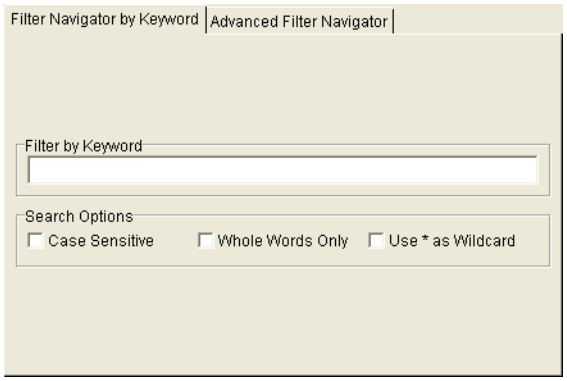


Figure 7-27 Filter Navigator Window: Filter by Keyword Tab

Filter by Keyword

The Filter by Keyword field lets you perform Boolean searches using the terms, “NOT”, “OR”, or “AND”. The Filter by Keyword field performs a text match on the following annotations:

- Name
- Folder Name
- Author/Owner/Uploader
- Research Group or Organization

- Notes
- Attachments
- Experiment

Search Options

The Search Options box contains the following options:

- **Case Sensitive**—Searches only for words using the specific capitalization you entered.
- **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
- **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.

Search Within Projects

The Search Within Projects box lists all projects for which the currently opened genome contains one or more data objects that are assigned to this project. If the genome does not contain any objects that are assigned to any project, only the special “Not Assigned” project is shown. This special project contains all those objects that are not assigned to any project. It lets you limit a search by project assignment and search for data objects assigned to one or more projects.

- **Check All**—Selects all of the projects.
- **Clear All**—Clears the selected projects.

Advanced Filter Navigator Tab

The Advanced Filter Navigator tab (Figure 7-28) provides more options for filtering your search. The search criteria adhere to the following requirements:

- All of the entries accept standard Boolean constructions, except the Creation Date fields.
- A blank field is equivalent to searching for anything (for that field)

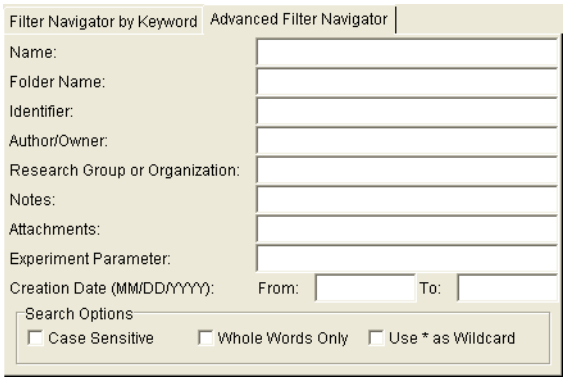


Figure 7-28 Filter Navigator Window: Advanced Filter Navigator Tab

Search Fields

- **Name**—Searches the names of the data objects.
- **Folder Name**—Searches the names of folders containing the selected the data objects.
- **Author/Owner**—Searches the Authors/Owners of the data objects.
- **Research Group or Organization**—Searches the names of the research group or organization.
- **Notes**—Only searches the notes associated with a data object. Only notes for object stored locally are searched. You cannot search for notes stored on Signet using the Filter Navigator.

  To find objects that are stored on Signet based on terms in the notes section, use the **Edit > Search** functionality. Go to "Searching for Data Locally or Remotely" on page 7-6 for more information.
- **Attachments**—Searches for attachments with a given name. It does not search the contents of attached files.
- **Experiment Parameters**—Limits the search to experiment based on either parameter names or the values of the parameter.
- **Creation Date**—Searches by date using MM/DD/YYYY format. If either of the Creation Date boxes are left blank it equates to "anything." For example: "from 6/13/03 to today" or "from any day to 6/13/03".

Search Options

The Search Options box contains the following options:

- **Case Sensitive**—Searches only for words using the specific capitalization you entered.
- **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string "statistic", the search results will contain samples that contain the word "statistic" in the specified fields, but ignore those containing "statistic" as part of a larger word, such as "statistician".

- **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.

Show Menu

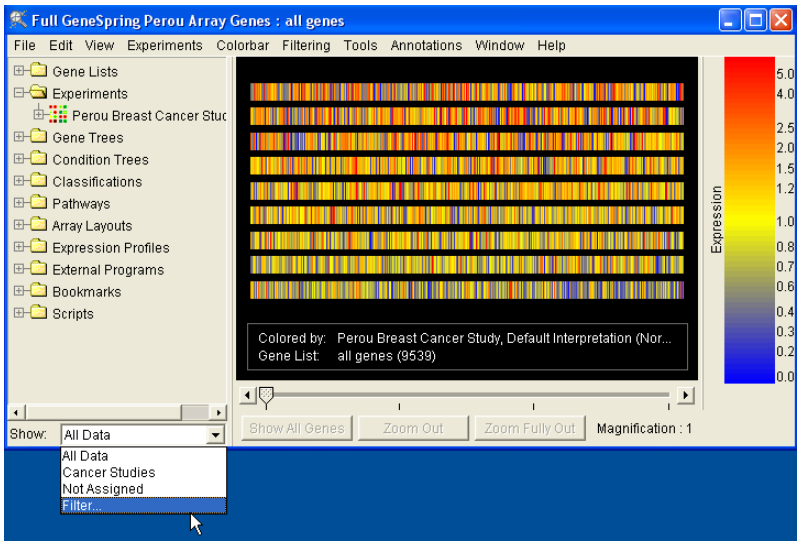
The Show menu, located at the bottom of the Navigator, also lets you filter data by assigned projects or unassigned projects. This menu contains the following commands:

- **All Data—**Shows all data objects in this genome.
- **Not Assigned—**Shows all data objects in the genome that have not been assigned to any project.
- **<*Project name*>—**Shows all data objects in the genome that have been assigned to the selected project.
- **Filter…**—Opens the Filter Navigator window.

Opening the Filter Navigator Window

To open the Filter Navigator window:

- Do one of the following:
  - In the **Navigator**, select **Filter** from the **Show** menu.



  - Select **Edit > Filter Navigator**.

Assigning Data Objects to Projects

You assign data objects to projects using the Change Projects window (Figure 7-29). The following options are available for accessing this window:

- Use the Search Results window
- Use a folder
- Use an Inspector window
- Use the Sample Manager

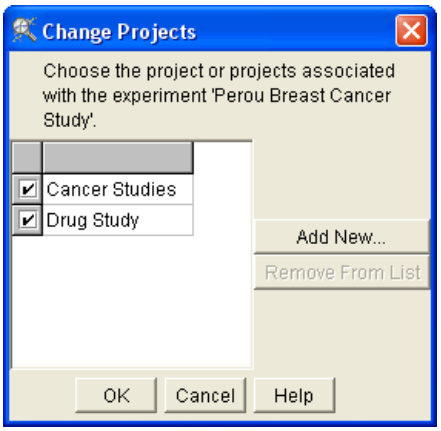


Figure 7-29 Change Projects Window

The Change Projects window contains a list of all project names that GeneSpring knows about, including projects that contain data objects that are stored on Signet.

You use this window to change the project assignment for all of the selected objects. The check boxes overwrite the existing project assignment. A partially-selected check box indicates that some, but not all, of the selected objects have been assigned to that project assignment.

Selecting a check box will add that project to all of the selected objects.Clearing the check box will remove that project from all of the selected objects.

The following section explains how to existing objects to projects using these options.

Using the Search Results Window

To use the Search Results window to assign data objects to projects:

1. Select **Edit > Search**.

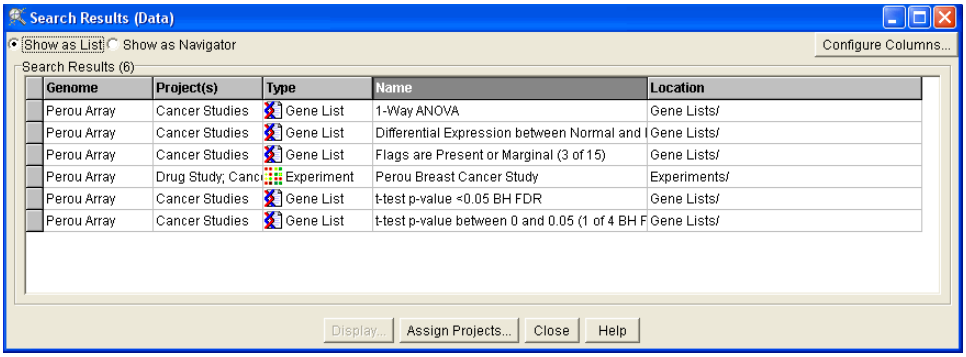
   The Search window opens.

2. Enter your search criteria.

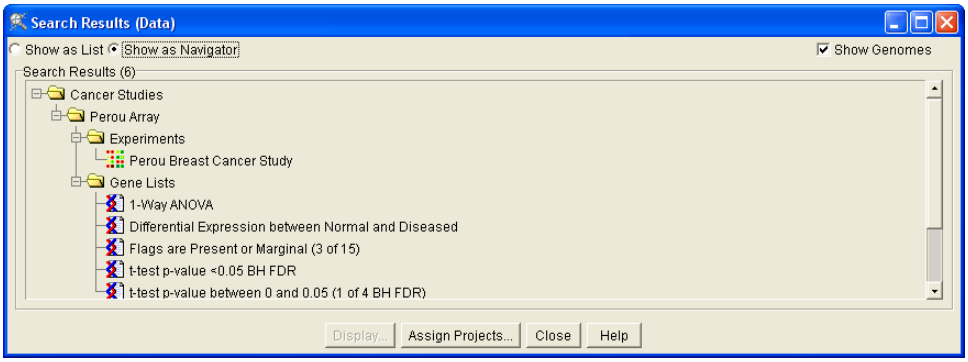
   Go to "Searching for Data Locally or Remotely" on page 7-6 for complete information on this window.

3. Click the **Start** button.

When the search is complete, a list of the data objects that match the search criteria appears in the Search Results window opens in the **Show as List** view.



4. To show the results in the Navigator, select the **Show as Navigator** option.



5. To assign all the data objects to one or more projects, do the following:
   a. Click the **Assign Projects** button.

      The Change Projects window opens.
   b. Click the **Add New** button.

      The Add New window opens.
   c. In the **Project Name** field, enter a name for the project.
   d. Click the **OK** button.
6. If the search results contain data objects that you do not want to assign to a project, do the following:
   a. Select the data object you want to remove and click the **Remove from List** button.
   b. Click the **OK** button.

7. To assign a previously created project name, do the following:
   a. Select the project name(s) you want.

      You can assign data objects to more than one project.

   b. Click the **OK** button.

Using Folders

You can assign all the objects in a folder to one or more projects all at once. This procedures lets you quickly assign data objects to a project.

To use a folder to assign data objects to projects:

1. In the **Navigator**, locate the folder that contains the data objects you want to add to a project.
2. Right-click the folder and select the **Assign Projects** command.

   The Change Projects window opens.
3. To create a new project name, do the following:
   a. Click the **Add New** button.

      The Add New window opens.

   b. In the **Project Name** field, enter a name for the project.

      All data objects in the project will be associated with this project name.

   c. Click the **OK** button.
4. To assign a previously created project name, do the following:
   a. Select the project name(s) you want.

      You can assign data objects to more than one project.

   b. Click the **OK** button.

Using Inspectors

To use an Inspector to assign data objects to projects:

1. Open the **Inspector** you want.

   Go to "Using Inspectors" on page 7-15 for more information.
2. Click the **Change Projects** button.

   The Change Projects window opens.
3. To create a new project name, do the following:
   a. Click the **Add New** button.

      The Add New window opens.

   b. In the **Project Name** field, enter a name for the project.

   c. Click the **OK** button.

4. To assign a previously created project name, do the following:
   a. Select the project name(s) you want.

      You can assign data objects to more than one project.

   b. Click the **OK** button.

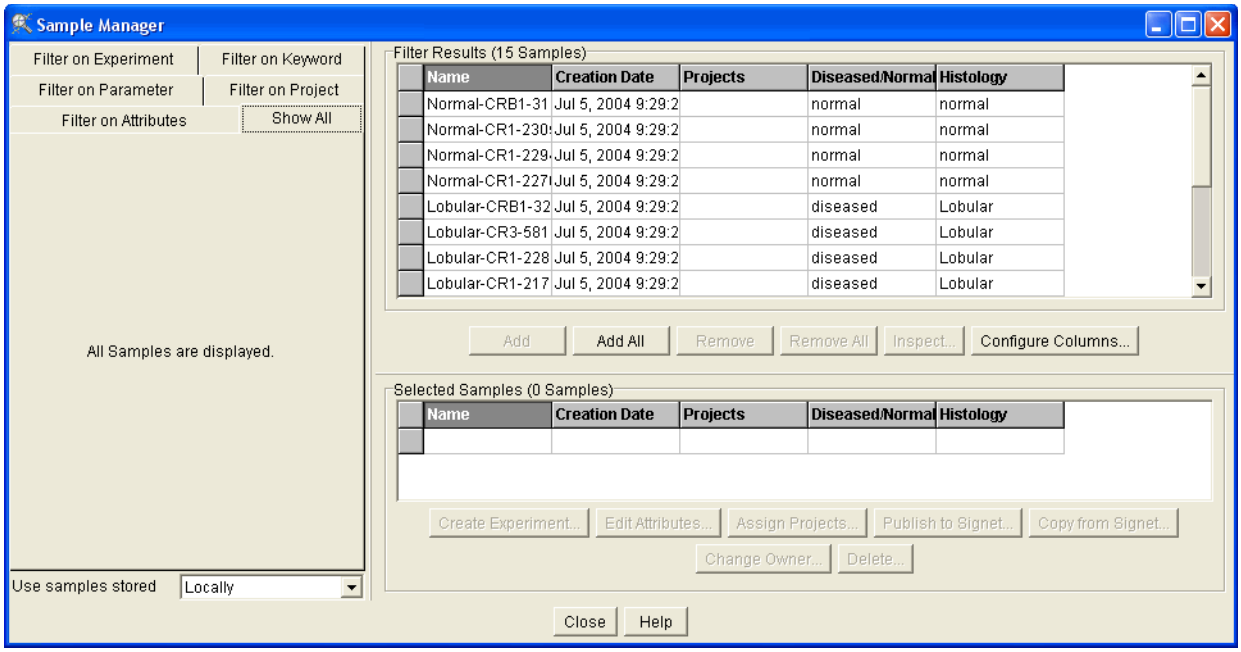
Using the Sample Manager

This section explains how to use the Sample Manager to assign data objects to projects. For more information about the Sample Manager, go to "Using the Sample Inspector" on page 7-19.

To use the Sample Manager to assign data objects to projects:

1. Select **Experiments > Sample Manager**.

   The Sample Manager window opens.



2. In the **Filter Results** column, locate the sample you want.
3. Click the **Assign Projects** button.

   The Change Projects window opens.

4. To create a new project name, do the following:
   a. Click the **Add New** button.

   The Add New window opens.

   b. In the **Project Name** field, enter a name for the project.
   c. Click the **OK** button.
5. To assign a previously created project name, do the following:

a. Select the project name(s) you want.

   You can assign data objects to more than one project.

b. Click the **OK** button.

Assigning New Experiments to Projects

To assign a new experiment to a project:

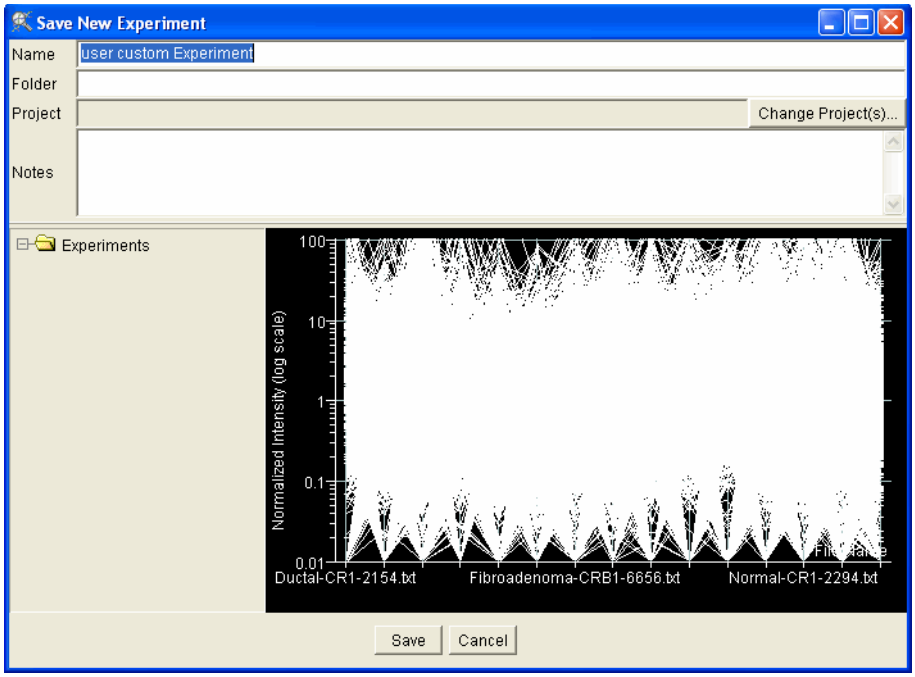
1. Select **Experiments > Create New Experiment**.

   The Create New Experiment window opens.

2. Create a new experiment.

   Go to "Creating New Experiments" on page 5-25 for information.

   After you create an experiment, the Save New Experiment window opens.



3. In the **Name** field, enter a name for the experiment. Be sure to choose a descriptive name that you will remember later.

4. To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.

5. To save in a new subfolder, navigate to the parent folder and enter a name for the new folder in the **Folder** field.

6. In the **Notes** field, enter any descriptive information about the experiment you want.

7. To save the experiment with a project, click the **Change Projects** button.

   The Change Projects window opens.

8. To create a new project name, do the following:
   a. Click the **Add New** button.

      The Add New window opens.
   b. In the **Project Name** field, enter a name for the project.
   c. Click the **OK** button.
9. To assign a previously created project name, do the following:
   a. Select the project name(s) you want.

      You can assign data objects to more than one project.
   b. Click the **OK** button.
10. Click the **Save** button.

8

Viewing Data

This chapter explains the different options that are available to view genes or associated data objects. The following topics are covered:

- Managing Window Elements
- Displaying Expression Data
- Changing Common Display Options

Managing Window Elements

GeneSpring offers various options for managing window elements. This section explains how to use these options. It covers the following topics:

- Linking Windows
- Splitting Windows
- Using Bookmarks to Save Your Place
- Showing or Hiding Window Display Elements

Linking Windows

The Linked windows function lets you select one gene or gene list in two windows simultaneously. Simply select a gene or gene list in one window and the same gene or gene list is automatically selected in the other window.

To create a linked window:

- Select **File** > **New Linked Window**.

Splitting Windows

Another interesting way to view classifications is with the Split windows function. The Split windows feature lets you see multiple sets simultaneously in the main GeneSpring window.

Figure 8-1 shows an example split window with a k-means clustering analysis colored by expression values. Note the list name and number of genes shown in the upper right corner of each small window. In this instance, the names are set numbers from the original k-means clustering.

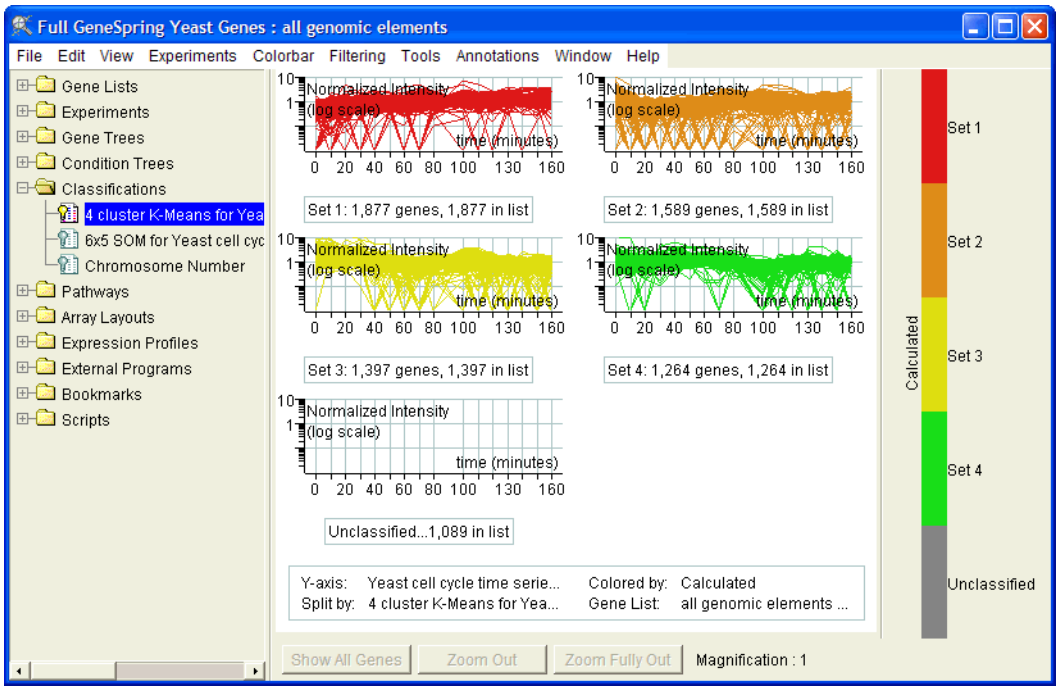


Figure 8-1 Example Split Window

To split a window:

1. In the **Navigator**, open the **Classifications** folder.
2. Right-click a classification and select the **Split Window** command.
3. Select one of the options:
   - Horizontally
   - Vertically
   - Both
   - Neither

   The main window of the Genome Browser splits into several small windows. Notice the number of genes beneath each small window. In addition, clicking any classification automatically splits the window in both directions.

   Note: In the Eisen-like subtree view, the thumbnail of the full tree remains in its usual position, but no marquee is shown specifying the subtree that has been zoomed in on. Each classification shows the same subtree.

To unsplit a window:

- Do one of the following:
  - Select **View > Unsplit window**.
  - Right-click over the original data object and select **Split > Neither**.

Using Bookmarks to Save Your Place

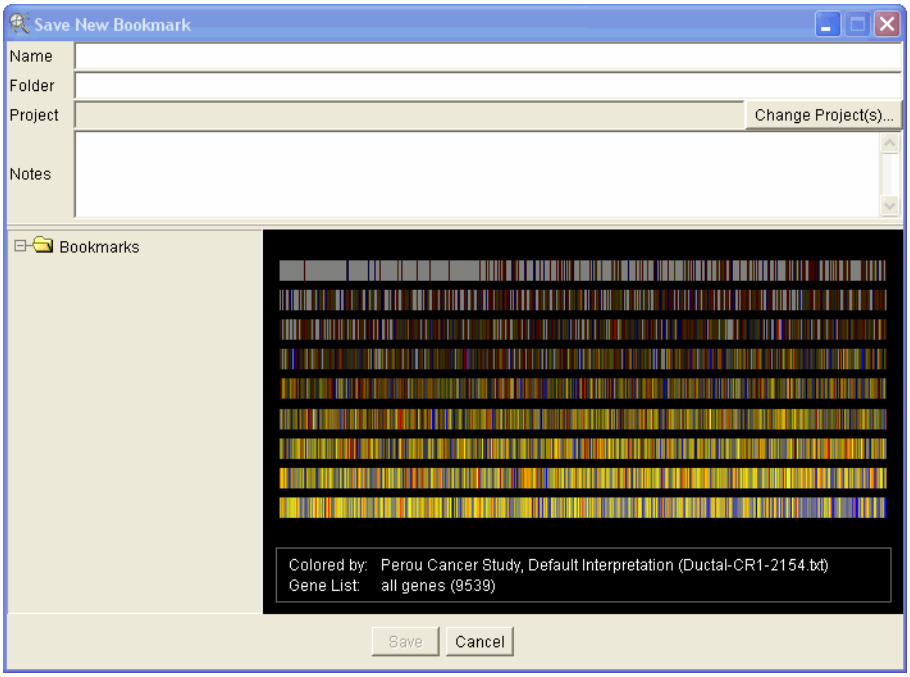
If you ever need to pause in the midst of your analysis, you can create a bookmark to hold your place. The **Bookmark** saves all your current display settings, including experiment, gene list, coloration, and selected genes.

Creating Bookmarks

To create a bookmark:

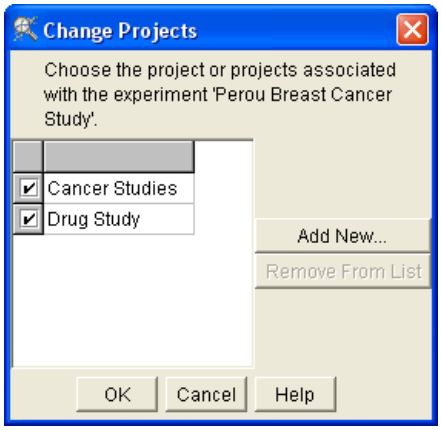
1. Select **File** > **Save Bookmark**.

   The Save Bookmark window opens.



2. In the **Name** field, enter a name for your bookmark.
3. In the **Folder** field, specify a folder in which to save the bookmark.
4. In the **Notes** field, add any notes you would like.
5. To save the bookmark in a project, click the **Change Projects** button.

   The Change Projects window opens.



6. Do one of the following:
   - To create a new project name, do the following:
     a. Click the **Add New** button.

        The Add New window opens.

     b. In the **Project Name** field, enter a name for the project.

        All data objects in the project will be associated with this project name.

     c. Click the **OK** button.
   - To assign a previously created project name, do the following:
     a. Select the project name(s) you want.

        You can assign data objects to more than one project.

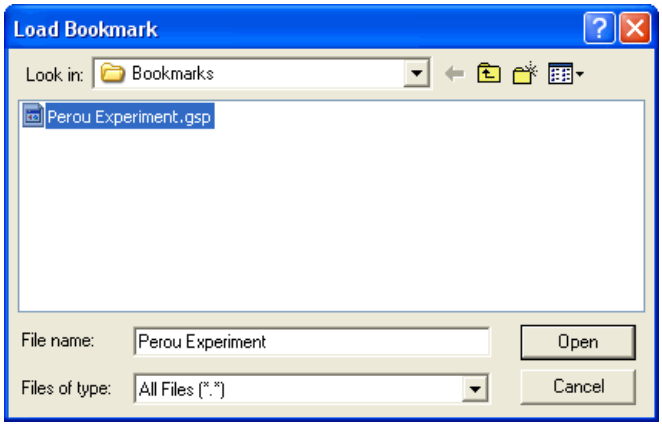
     b. Click the **OK** button.
7. Click **Save**.

Loading Bookmarks

To load a bookmark:

1. Do one of the following:
   - In the **Navigator**, open the **Bookmarks** folder, and then click the bookmark.
   - Select **File** > **Load Bookmark File**.

   The Load Bookmark window opens.



2. Select the bookmark you want and click **Open**.

Showing or Hiding Window Display Elements

You have the option of showing or hiding many of the elements in the GeneSpring window.

To show or hide window display elements:

1. Select **View > Visible**.
2. Select one of the following options:
   - **Picture—**Shows or hides the optional picture at the bottom right corner of the window.
   - **Animation Controls—**Shows or hides the slider and the Animate check box at the bottom of the window (hiding this check box does not disable the Animation feature).
   - **Magnification—**Shows or hides the Magnification feature and the **Zoom Out** button at the bottom of the window (hiding the **Zoom Out** button does not disable the **Zoom Out** menu option).
   - **Secondary Picture—**Shows or hides your secondary picture when you are viewing two gene lists or experiments simultaneously in the Genome Browser.
   - **Secondary Animation Controls—**Shows or hides the secondary Animation Controls check box and slider when you are viewing two gene lists or experiments simultaneously.
   - **Navigator—**Shows or hides the Navigator panel.
   - **Hide All—**Hides everything in the window except the Genome Browser.
   - **Show All—**Shows all elements.
   - **Hide All in All Windows—**Hides everything in all windows except the Genome Browser.
   - **Show All in All Windows—**Shows all elements in all windows.

Displaying Expression Data

GeneSpring displays expression data in a manner that will help you conceptualize important results and enhance your publications. Data can be viewed with highly-customizable visualization tools including:

- 2D and 3D scatter plots
- 2D dendrograms
- Chromosome maps
- Pathway diagrams
- Venn diagrams
- Classification views

The GeneSpring tools for viewing your genomic and array data are provided in the View menu in the main GeneSpring window. This section covers the following topics:

- Using the Blocks View
- Using the Graph View
- Using Bar Graph View
- Using Physical Position View
- Using the Scatter Plot View
- Using the 3D Scatter Plot View
- Using the Tree View
- Using the Ordered List View
- Using the Array Layout View
- Using the Pathway View
- Using the Compare Genes to Genes View
- Using the Graph by Genes View
- Using the Spreadsheet View
- Using the Condition Scatter Plot

Using the Blocks View

The Blocks view (Figure 8-2) displays a rectangle for every gene in the active genome, ordered by trust.

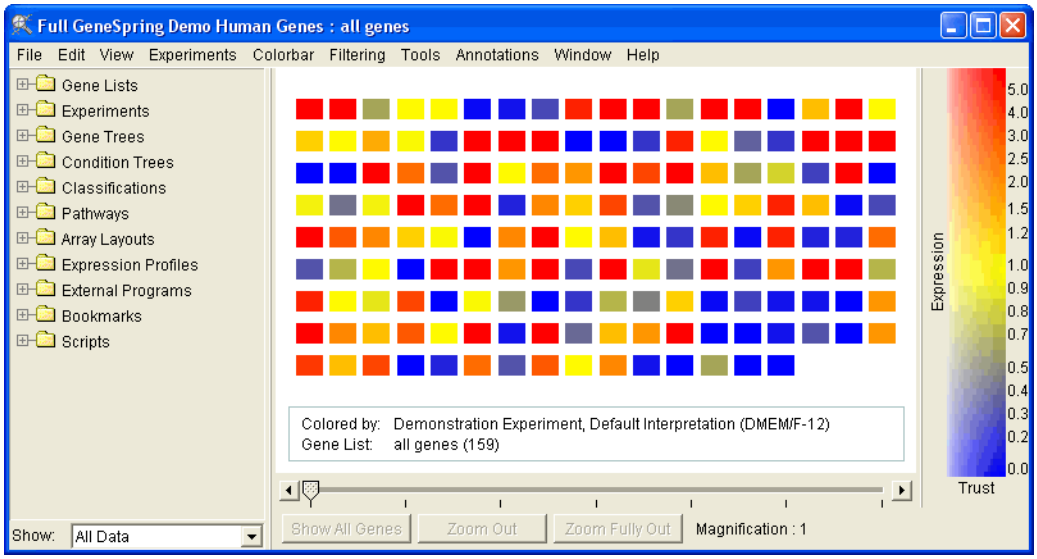


Figure 8-2 Blocks View

Opening the Blocks View

To open the Blocks view:

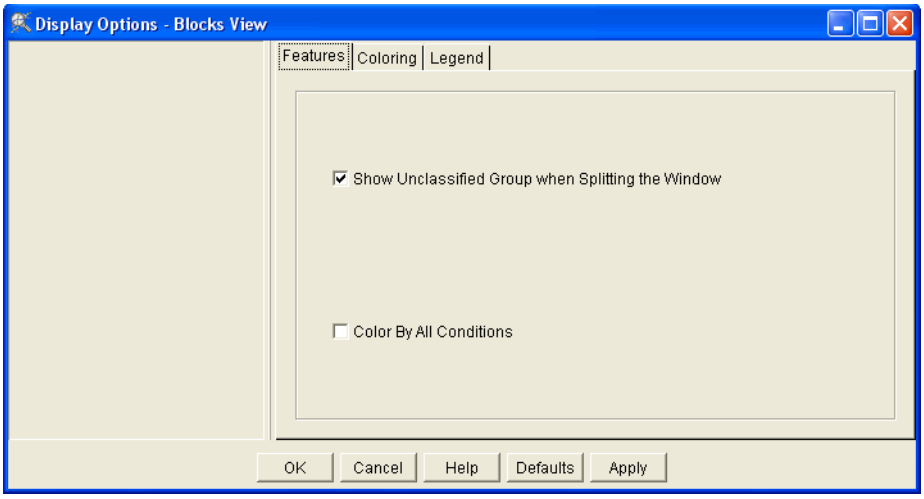
- Select **View > Blocks**.

Displaying Options for the Blocks View

To display options:

- Do one of the following:
  - Open the **Blocks** view and then select **View > Display Options**.
  - Open the **Blocks** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

The Display Options window opens.



Changing Features

The Features tab contains a column of check boxes that allow you toggle on or off certain items in the Genome Browser.

To change features:

1. Open the **Display Options** window for the Block view.

   Go to "Displaying Options for the Blocks View" on page 8-7 for information.

2. Click the **Features** tab.

3. Select the features you want:

   - **Color by all conditions**—Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the Genome Browser.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

Changing Color Settings

See "Changing Display Options for Color Settings" on page 8-60.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Using the Graph View

The Graph view (Figure 8-3) lets you visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a line. Genes with no data cannot be displayed in this view.

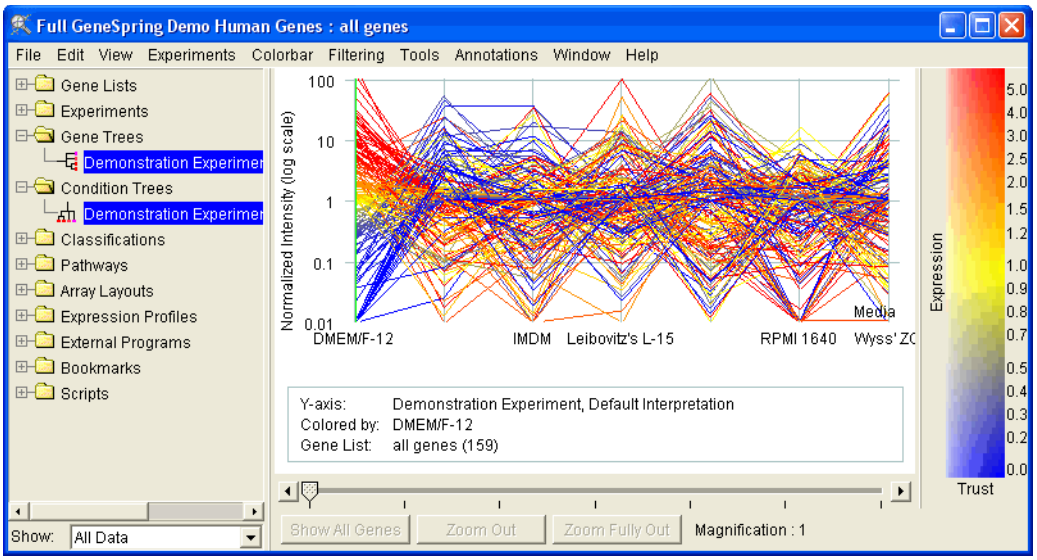


Figure 8-3 Graph View

Opening the Graph View

To open the Graph view:

- Select **View** > **Graph**.

  The Graph option consists of two views: the continuous graph view and the histogram view, which appears if the experiment being displayed contains any non-continuous parameters.

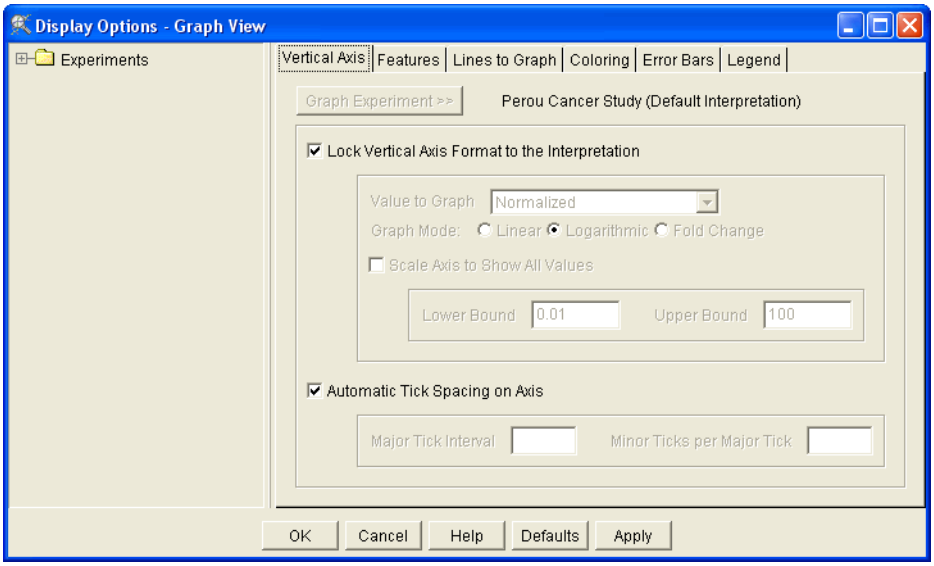
  The figure above shows the genes in the “all genes” list in Graph view. The gene in white has been selected; its name appears in the legend, after the name of the gene list.

Displaying Options for the Graph View

To display options:

- Do one of the following:
  - Open the **Graph** view and then select **View** > **Display Options**.
  - Open the **Graph** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

The Display Options window opens.



Changing the Vertical Axis

To change the vertical axis:

1. Open the **Display Options** window for the Graph view.

   Go to "Displaying Options for the Graph View" on page 8-9 for information.

   The Display Options window opens.
2. Click the **Vertical Axis** tab.
3. Clear the **Lock Vertical Axis Format to the Interpretation** check box.
4. In the **Value to Graph** list, select one of the following options:
   - Normalized
   - Control
   - Raw
   - Average of Raw and Control
   - Max of Raw and Control.
5. In the Graph Mode box, select one of the following options:
   - Linear
   - Logarithmic
   - Fold Change
6. To adjust the vertical axis so that all measurements are visible, check the **Scale Vertical Axis to Show all Values** box.

   The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.

7. To adjust tick spacing, do the following:
   a. Clear the **Automatic Tick Spacing on Vertical Axis** box.
   b. Enter the distance between major ticks in the **Major Tick Interval** field.
   c. Enter the number of divisions between major ticks in the **Minor Ticks per Major Tick** field.

      Note: The number of visible tick-marks between major ticks is one less than the number you enter.
8. Click **Apply**.

Changing Features

The Features tab contains a column of check boxes that allow you toggle on or off certain items in the Genome Browser.

1. Open the **Display Options** window for the Graph view.

   Go to "Displaying Options for the Graph View" on page 8-9 for information.
2. Click the **Features** tab.
3. Select the features you want:
   - **Show Experiment Name—**Displays the name of the current experiment in the upper right-hand corner of the Genome Browser.
   - **Show Horizontal Axis Label—**Displays the parameter that is graphed on the horizontal axis.
   - **Show Vertical Axis Label—**Displays the parameter that is graphed on the vertical axis.
   - **Label Vertical Axis on Side—**Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
   - **Show Condition Line—**Displays the vertical bar that can be moved with the condition slider.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.
4. Click **Apply**.

Changing Lines to Graph

You have the option to draw grid lines to help distinguish distinct groups of data points.

To change lines:

1. Open the **Display Options** window for the Graph view.

   Go to "Displaying Options for the Graph View" on page 8-9 for instructions.
2. Click the **Lines to Graph** tab.

To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis.

3. Select any of the **Major Tick Intervals/Minor Tick Intervals** check boxes you want.
4. Click **Apply** to view your data with grid lines.

Changing Color Settings

See "Changing Display Options for Color Settings" on page 8-60.

Using Error Bars

Error Bar Information displays whether the error bar is based on Standard Error, Standard Deviation or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Using Bar Graph View

The Bar Graph view (Figure 8-4) lets you visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a vertical bar. Genes with no data cannot be displayed in this view.

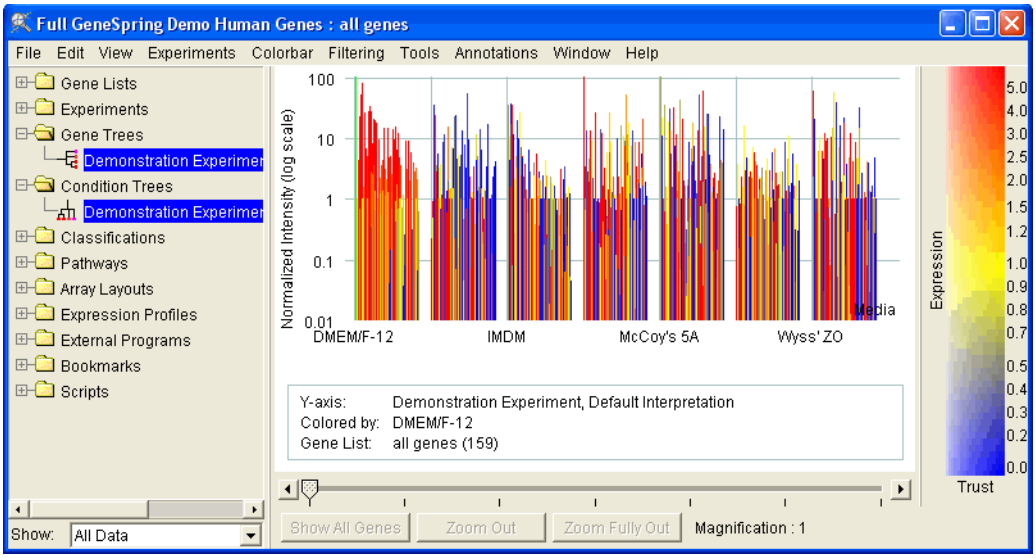


Figure 8-4 Bar Graph View

Opening the Bar Graph View

To open the Bar Graph view:

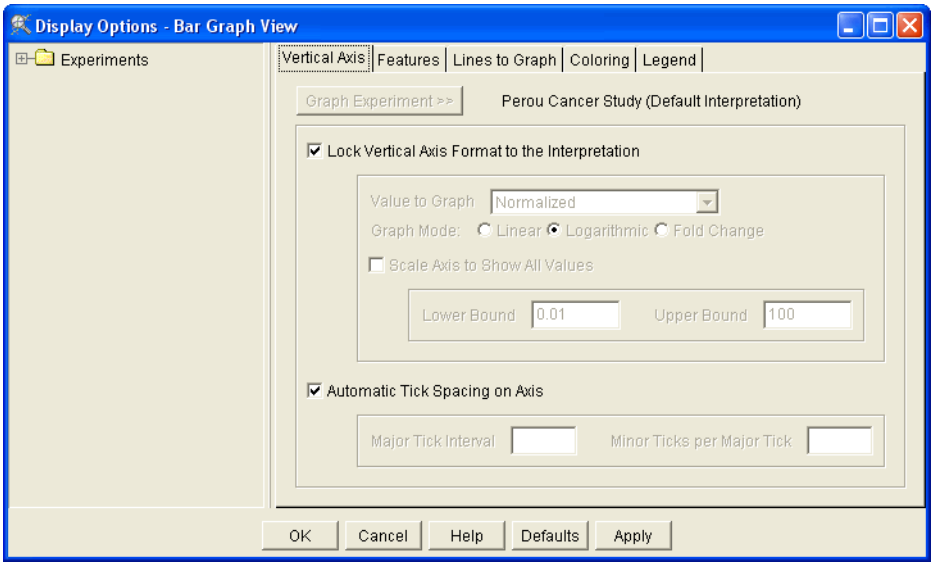
- Select **View** > **Bar Graph**.

Displaying Options for the Bar Graph View

To display options:

- Do one of the following:
  - Open the **Bar Graph** view and then select **View** > **Display Options**.
  - Open the **Bar Graph** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

    The Display Options window opens.



Changing the Vertical Axis

To change the vertical axis:

1. Open the **Display Options** window for the Bar Graph view.

   Go to "Displaying Options for the Bar Graph View" on page 8-13 for information.

   The Display Options window opens.
2. Click the **Vertical Axis** tab.
3. Clear the **Lock Vertical Axis Format to the Interpretation** check box.
4. In the **Value to Graph** list, select one of the following options:
   - Normalized
   - Control
   - Raw

- Average of Raw and Control
- Max of Raw and Control

5. In the **Graph Mode** box, select one of the following options:
   - Linear
   - Logarithmic
   - Fold Change
6. To adjust the vertical axis so that all measurements are visible, check the **Scale Vertical Axis to Show all Values** box.

   The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.
7. To adjust tick spacing, do the following:
   a. Clear the **Automatic Tick Spacing on Vertical Axis** box.
   b. Enter the distance between major ticks in the **Major Tick Interval** field.
   c. Enter the number of divisions between major ticks in the **Minor Ticks per Major Tick** field.

      Note that the number of visible tick-marks between major ticks is one less than the number you enter.
8. Click **Apply**.

Changing Features

The Features tab contains a column of check boxes that lets you toggle on or off certain items in the Genome Browser.

To change features:

1. Open the **Display Options** window for the Bar Graph view.

   Go to "Displaying Options for the Bar Graph View" on page 8-13 for information.
2. Click the **Features** tab.
3. Select the features you want:
   - **Show Horizontal Axis Label—**Displays the parameter that is graphed on the horizontal axis.
   - **Show Vertical Axis Label—**Displays the parameter that is graphed on the vertical axis.
   - **Label Vertical Axis on Side—**Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
   - **Show Condition Line—**Displays the vertical bar that can be moved with the condition slider.
   - **3D Look—**Places the bars on a diagonal line so as to imply that genes in each condition are stacked in rows perpendicular to the horizontal axis.

- **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

4. Click **Apply**.

Changing Lines to Graph

You have the option to draw grid lines to help distinguish distinct groups of data points.

To change lines:

1. Open the **Display Options** window for the Bar Graph view.

   Go to "Displaying Options for the Bar Graph View" on page 8-13 for information.

2. Click the **Lines to Graph** tab.

   To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis.

3. Select any of the **Major Tick Intervals/Minor Tick Intervals** check boxes you want.
4. Click **Apply** to view your data with grid lines.
5. Click **Apply**.

Changing Color Settings

See "Changing Display Options for Color Settings" on page 8-60.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Using Physical Position View

The Physical Position view lets you see an experiment or a set of experiments by organizing the genes according to their physical position (when the gene loci are known and loaded into GeneSpring) within the DNA sequence of the organism.

The Physical Position view works for any organism whose mapping data is at least partially available. An illustration of what Physical Position View looks like for humans is given in Figure 8-5.

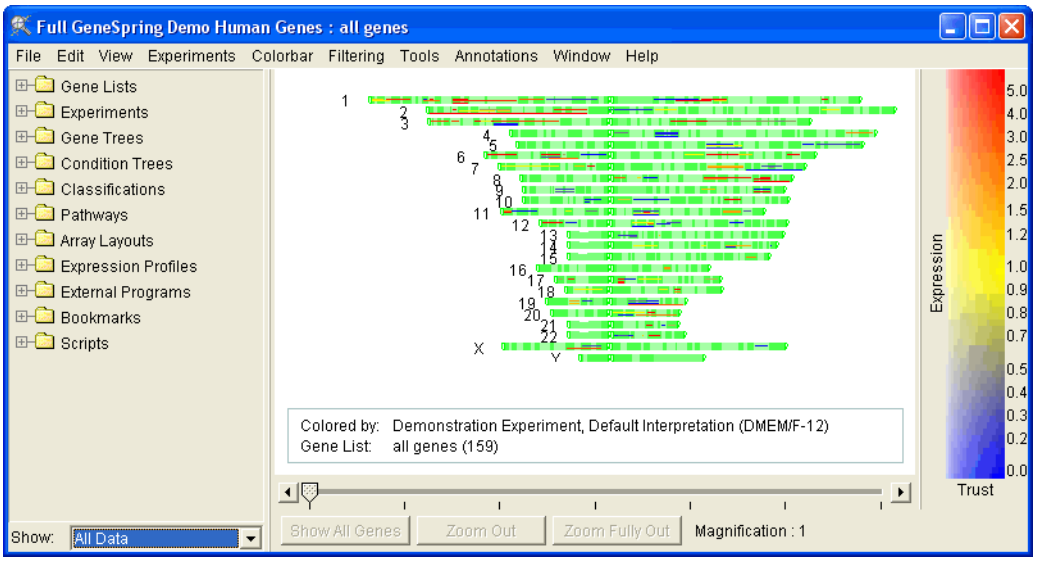


Figure 8-5 Physical Position View for a Human Genome

For organisms already sequenced, the physical position views looks more like yeast (illustrated in Figure 8-6).

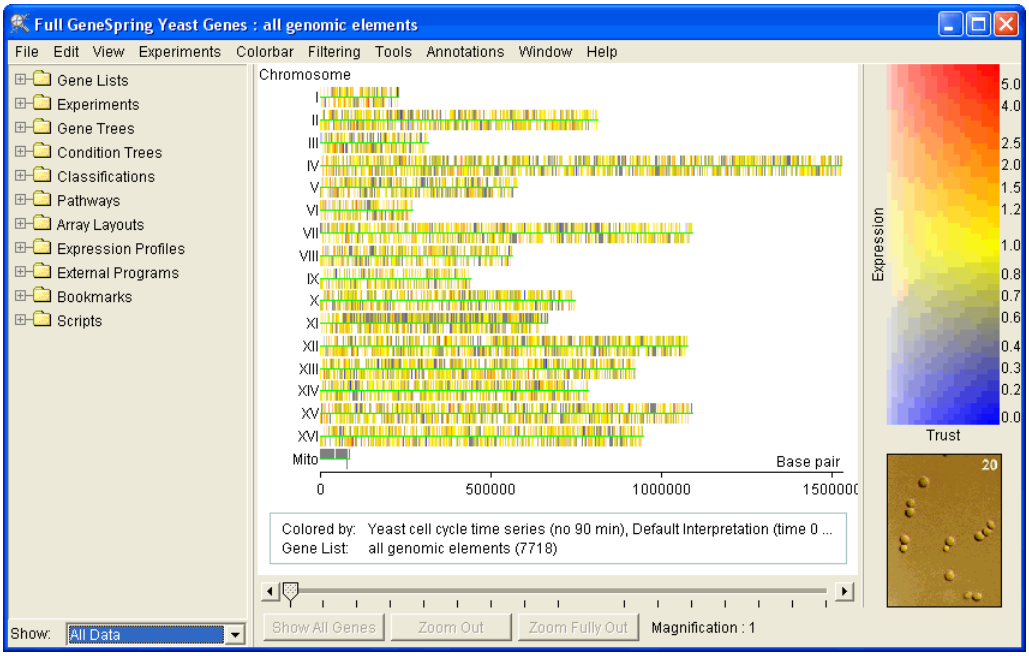


Figure 8-6 Physical Position View for Yeast Organisms

At greater magnification, you can see the base pairs (Figure 8-7).

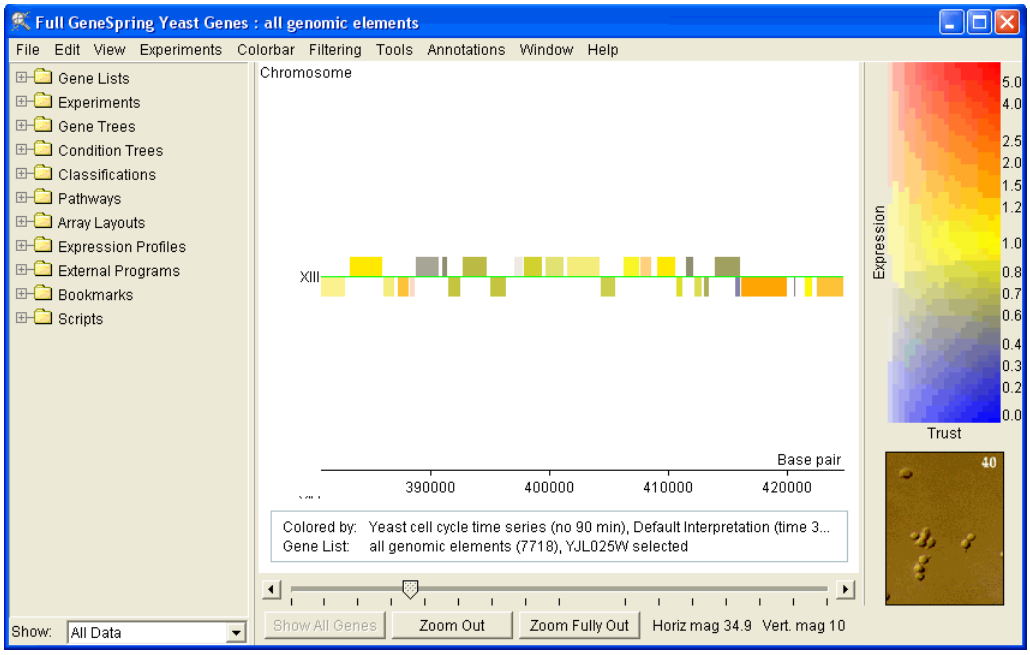


Figure 8-7 Base Pairs on Chromosome XIII

At greater magnification, the labels associated with the chromosome's cytogenetic bands are also visible.

Opening the Physical Position View

To open the Physical Position view:

- Select **View > Physical Position**.

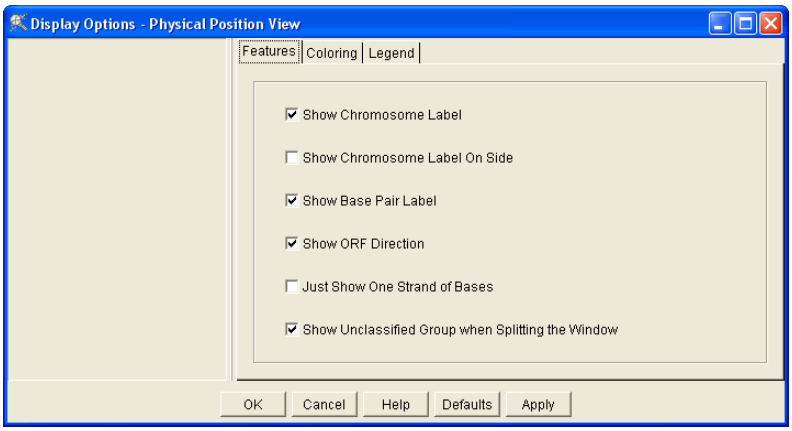
In GeneSpring versions 4.0 and later, sequence information is loaded by default if it is available. If you have an old version of GeneSpring and cannot update it, contact Silicon Genetics at 1-866-SIG-SOFT (744-7638).

Displaying Options for the Physical Position View

To display options:

- Do one of the following:
  - Open the **Physical Position** view and then select **View > Display Options**.
  - Open the **Physical Position** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

The Display Options window opens.



Changing Features

The Features panel of the display options window contains a column of check boxes that lets you toggle on or off certain items in the Genome Browser.

To change features:

1. Open the **Display Options** window for the Physical Position view.

   Go to "Displaying Options for the Physical Position View" on page 8-17 for information.

2. Click the **Features** tab.

3. Select the features you want:

   - **Show Chromosome Label—**Displays the word "Chromosome" next to the chromosome names or numbers.
   - **Show Chromosome Label on Side—**Displays the word "chromosome" vertically beside the chromosome names or numbers.
   - **Show Base Pair Label—**Displays the words "Base Pair" next to the axis representing the sequence location.
   - **Show ORF direction—**Places genes above or below the chromosomes depending on the direction they are transcribed. Genes on the top of the line are transcribed from left to right. Leaving this option unchecked places all of the genes on top of the chromosome lines.
   - **Show Just One Strand of Bases—**Displays only the bases on the Watson strand (when the Genome Browser is zoomed-in enough to display them).
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

4. Click **Apply**.

Changing Color Settings

See "Changing Display Options for Color Settings" on page 8-60.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57

Loading Sequences

The Load Sequence command is applicable only for sequenced organisms. Load the nucleic acid sequence to magnify a section of the physical position view until the nucleic acid sequence is displayed. Loading the sequence also lets you take advantage of GeneSpring's other sequence-based features such as **Tools > Find Potential Regulatory Sequences**. You can load the nucleic acid sequence in a number of ways.

To load a sequence that takes effect immediately:

1. Open the **Physical Position** view.

   Go to "Opening the Physical Position View" on page 8-17 for instructions.

2. Right-click in the Genome Browser and select **Options > Load Sequence**.

   A window saying *Please wait while nucleic acid sequence is loaded* appears. After the loading is complete it is possible to zoom in and see the nucleic acid sequence of a particular gene.

   The sequence is shown in the magnified genes. However, this information is not saved, so when you exit and re-open GeneSpring you must reload the nucleic acid sequence.

   If you would like the sequences to always be readily available, you must change the defaults through the Preferences window. You may choose to make the load sequence feature automatically load with the program. Again, note that this applies to version 4.0 and earlier.

To load a sequence that takes effect at your next GeneSpring session:

1. Open the **Physical Position** view.

   Go to "Opening the Physical Position View" on page 8-17 for instructions.

2. Select **Edit > Preferences**.

   The GeneSpring Preferences window opens.

3. Select **Data Files** from the pull-down at the top of the window.

4. Select the **Load Sequence** check box.

5. Click **OK** at the bottom of the window.

6. Do one of the following:

   - Close and restart GeneSpring.
   - Select **File > New Window**.

   Changing the defaults in the Preferences window does not initiate the load sequence feature in your current session, but it does change future initial loading practices. The

nucleic acid sequence can also be loaded as a side effect of using **Tools > Find Regulatory Sequences**. For more information on this particular feature, see "Finding Potential Regulatory Sequences" on page 14-1.

Mouse Cytogenetic Band Graphics

The physical position view is now able to show mouse genes mapped to their chromosomal locations. Simply drag and drop a zip file onto any mouse genome to add mouse cytogenetic band graphics. Go to http://www.silicongenetics.com/cgi/SiG.cgi/Support/resources.smf to get the zip file.

Using the Scatter Plot View

The Scatter Plot view (Figure 8-8) is useful for examining the expression levels of genes in two distinct conditions, samples, or normalization schemes. It is the most flexible of all the views in its ability to customize the way data are displayed.

For instance, you can use the scatter plot to identify genes that are differentially expressed in one sample versus another.

A scatter plot can also be used to compare two values associated with genes in two gene lists. Such associated values might include the relative contribution of principal components as determined from principal components analysis, or two similarity scores from the Find Similar function in the Gene Inspector. Genes with no data cannot be displayed in this view.

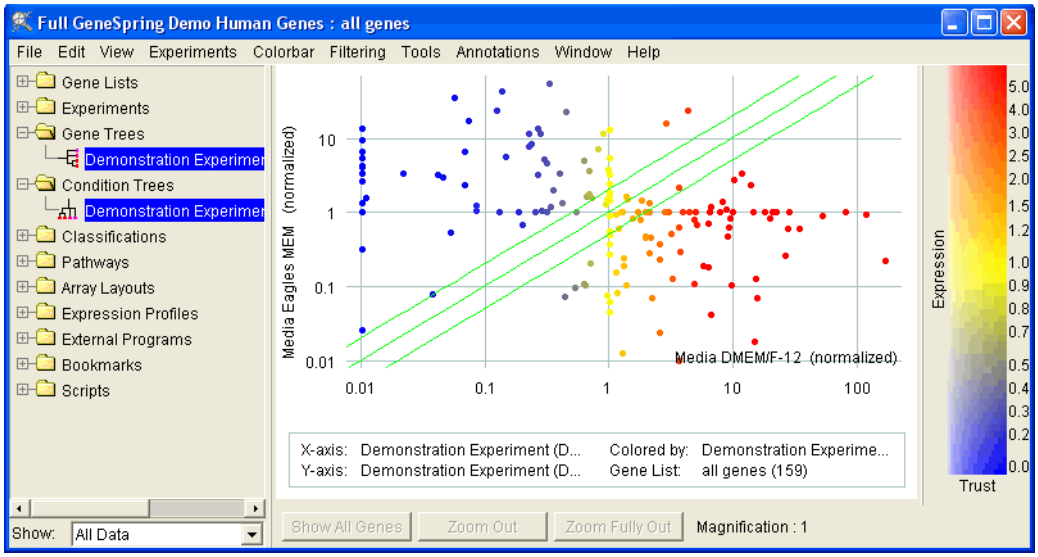


Figure 8-8 Scatter Plot View

In Figure 8-8, each '+' symbol represents a gene. The vertical position of each gene represents its expression level in the current condition, and the horizontal position represents its control strength (in this case, the median expression level of this gene in all conditions). Genes that fall above the diagonal are overexpressed and genes that fall below the diagonal are underexpressed as compared to their median expression level over the course of the experiment.

Opening the Scatter Plot View

To open the Scatter Plot view:

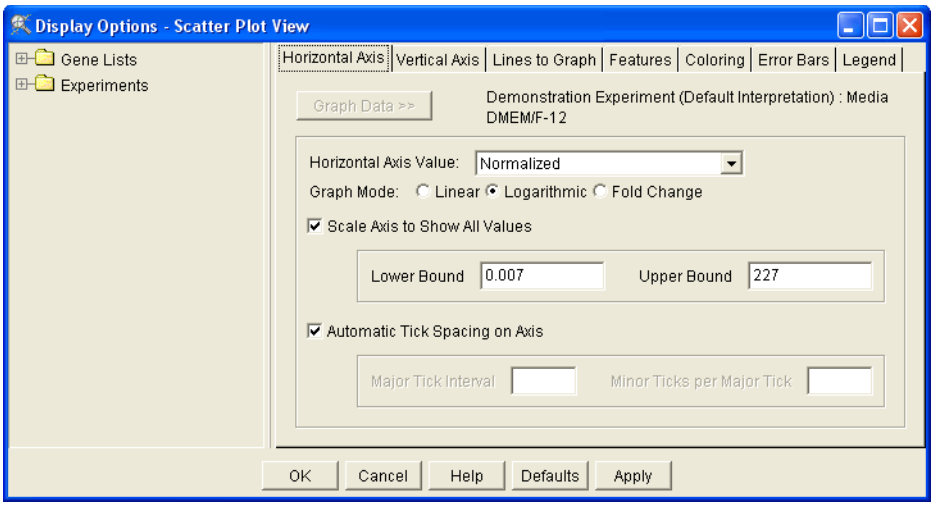
- Select **View** > **Scatter Plot**.

Displaying Options for the Scatter Plot View

To display options:

- Do one of the following:
  - Open the **Scatter Plot** view and then select **View** > **Display Options**.
  - Open the **Scatter Plot** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

    The Display Options window opens.



Changing the Vertical or Horizontal Axis

The most critical option to set in the Scatter Plot view is the type of data that is displayed on the two axes.

To change the vertical or horizontal axis:

1. Open the **Display Options** window for the Scatter Plot view.

   Go to "Displaying Options for the Scatter Plot View" on page 8-21 for information.

   The Display Options window opens.

2. Click the **Horizontal Axis** or **Vertical Axis** tab.

3. In the **Navigator**, select the gene list, experiment, interpretation, or condition to use on the selected axis.

4. Click the **Horizontal/Vertical Axis Value** menu.

   The list of options includes only those that are appropriate for the type of data object you selected.

5. Choose a graph mode for the specified axis.

   The three options are linear, logarithmic, and fold change. Note that the fold change option is only available if you are looking at normalized data from an interpretation or a condition.

6. To adjust the vertical axes so that all measurements are visible, check the **Scale Axis to Show all Values** box.

   The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.

7. Do one of the following:

   - To automatically choose tick spacings, select the **Automatic Tick Spacing on Axis** box.
   - To set the tick spacings manually, clear the check box and enter the major tick interval as well as the number of minor ticks.

   For more information about setting tick spacings, see "Changing the Vertical Axis" on page 8-13.

8. Click **Apply**.

Changing Features

The Scatter Plot view also lets you change the appearance of data points and data labels.

To change features:

1. Open the **Display Options** window for the Scatter Plot view.

   Go to "Displaying Options for the Scatter Plot View" on page 8-21 for information.

   The Display Options window opens.

2. Click the **Features** tab.

3. To modify the size and shape of the points, choose from the options you want in the **Style** and **Size** menus.

4. Select the options you want for the plot. The options are:

   - **Show Gene Names—**Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
   - **Show Horizontal Axis Label—**Displays the parameter that is graphed on the horizontal axis.
   - **Show Vertical Axis Label—**Displays the parameter that is graphed on the vertical axis.
   - **Label Vertical Axis on Side—**Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

5. Click **Apply**.

Changing Lines to Graph

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines.

These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the Genome Browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit.

To change lines:

1. Open the **Display Options** window for the Scatter Plot view.

   Go to "Displaying Options for the Scatter Plot View" on page 8-21 for information.

   The Display Options window opens.

2. Click the **Lines to Graph** tab.
3. To use fold change lines click the **Fold Change Lines** check box.
4. If you only want one pair of fold change lines, select the **Set Lines At option** and enter a number in the **fold** box.
5. If you would like more than one pair of lines, select the **Set Lines at Multiple Intervals** check box and list the "fold-values" to view, separated by commas.
6. To show a trend in your data, select the **Line of Best Fit** check box.

   Note that the regression is performed on the transformed data, and this line is always linear, regardless of how the axes are chosen.

7. To make the origin of each axis more visible, select the **Lines Through Origin** option.
8. To see a grid inside the plot area, select the **Horizontal/Vertical Grid Lines** check boxes.You can have lines drawn at the major and minor tick intervals of each axis.

   The color of these grid lines is represented in the Grid Color box at the bottom of the window.

9. To modify the grid color, click **Change**.
10. Click **Apply**.

Changing Color Settings

Coloring in the Scatter Plot view is more complicated than in other views because the color of each gene can be derived from the data in either axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot lets you color genes based on a third experiment or condition that is not plotted on *either* axis.

To change color settings:

1. Open the **Display Options** window for the Scatter Plot view.

   Go to "Displaying Options for the Scatter Plot View" on page 8-21 for information.

   The Display Options window opens.

2. Click the **Coloring** tab.

3. In **Color data points by** menu, select the type of data that is to be used for coloring.

   For more information about the types of data that are available for coloring, see "Changing Display Options for Color Settings" on page 8-60.

4. In **Use the expression levels in the following experiment or condition** options, select the axis you want to color.

   Note that only axes which represent experiments, interpretations, or conditions are available for coloring.

5. To color genes by an experiment that is not represented by either axis, do the following:

   a. Click **Other Experiment**.

   b. Select an experiment in the **Navigator**.

   c. Click **Set Experiment**.

6. Click **Apply**.

Using Error Bars

Error Bar Information displays whether the error bar is based on Standard Error, Standard Deviation or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57

Using the 3D Scatter Plot View

The 3D Scatter Plot view (Figure 8-9) lets you plot data on three axes. You can also show the centroid of each group when coloring by a classification. Genes with no data cannot be displayed in this view.

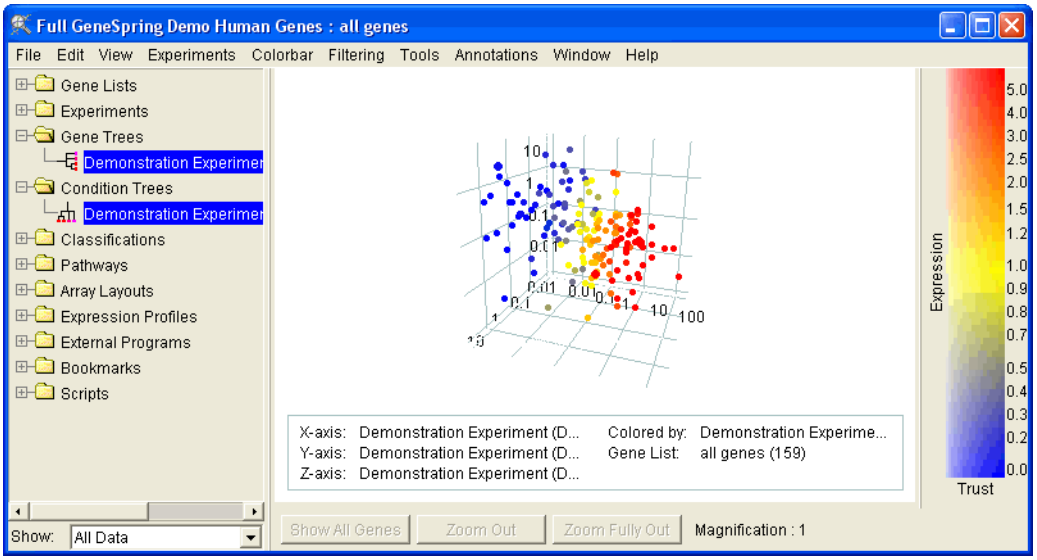


Figure 8-9 3D Scatter Plot View

In the 3D scatter plot above, each dot represents a gene. The vertical position of each gene represents its expression level in the current condition, and the horizontal position represents its control strength (in this case, the median expression level of this gene in all conditions).

Pressing the X, Y, or Z keys rotates the graph on the specified axis. Hold down the Shift key to speed this rotation. Hold down the Alt key to reverse the direction of rotation.

Opening the 3D Scatter Plot View

To open the 3D Scatter Plot view:

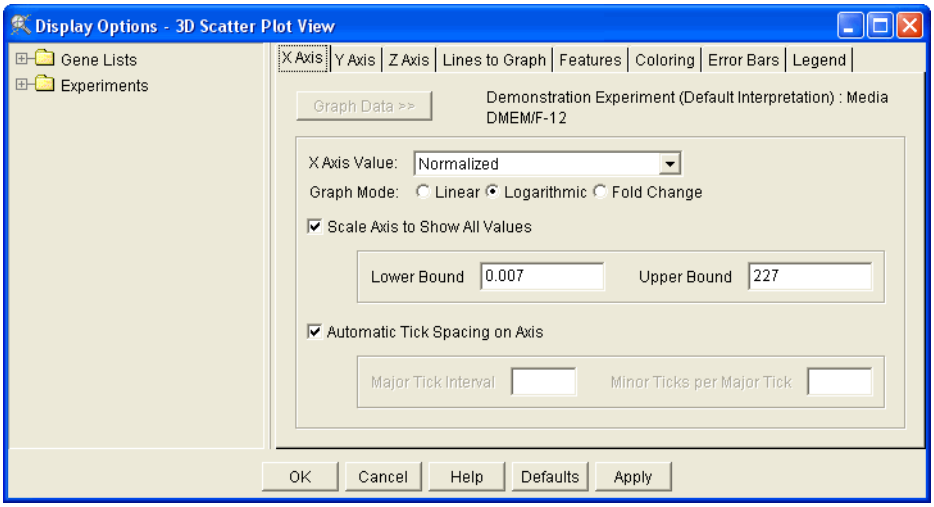
- Select **View > 3D Scatter Plot**.

Displaying Options for the 3D Scatter Plot View

To display options:

- Do one of the following:
  - Open the **3D Scatter Plot** view and then select **View > Display Options**.
  - Open the **3D Scatter Plot** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

The Display Options window opens.



Changing the X, Y, and Z Axes

The most critical option to set is the type of data that is displayed on the three axes.

To change the axes:

1. Open the **Display Options** window for the 3D Scatter Plot view.

   Go to "Displaying Options for the 3D Scatter Plot View" on page 8-25 for information.

   The Display Options window opens.

2. Click the **X Axis**, **Y Axis**, or **Z Axis** tab.

3. In the **Navigator**, select the gene list, experiment, interpretation, or condition to use on the selected axis.

4. Click the **X/Y/Z Axis Value** menu.

   The list of options includes only those that are appropriate for the type of data object you selected.

5. Choose a graph mode for the specified axis. The three options are linear, logarithmic, and fold change.

   Note that the fold change option is only available if you are looking at normalized data from an interpretation or a condition.

6. To adjust the axes so that all measurements are visible, do one of the following:

   - Select the **Scale Axis to Show all Values** check box.

     The upper and lower bounds are adjusted automatically.

   - Manually set the upper and lower bounds to values you want.

7. To choose tick spacings, do one of the following:
   - To automatically choose tick spacings, select the **Automatic Tick Spacing on Axis** box.
   - To set the tick spacings manually, leave this box unchecked and enter the major tick interval as well as the number of minor ticks.

   For more information about setting tick spacings, see "Changing the Vertical Axis" on page 8-13.
8. Click **Apply**.

Changing Lines to Graph

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines.

These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the Genome Browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit.

To change lines:

1. Open the **Display Options** window for the 3D Scatter Plot view.

   Go to "Displaying Options for the 3D Scatter Plot View" on page 8-25 for information.

   The Display Options window opens.
2. Click the **Lines to Graph** tab.
3. To see a grid inside the plot area, select the **Horizontal/Vertical Grid Lines** check boxes.You can have lines drawn at the major and minor tick intervals of each axis.

   The color of these grid lines is represented in the Grid Color box at the bottom of the window.
4. To modify the grid color, click **Change**.
5. Click **Apply**.

Changing Features

The 3D Scatter Plot view also lets you change the appearance of data points and data labels.

To change features:

1. Open the **Display Options** window for the 3D Scatter Plot view.

   Go to "Displaying Options for the 3D Scatter Plot View" on page 8-25 for information.

   The Display Options window opens.
2. Click the **Features** tab.
3. To modify the size and shape of the points, select the option you want in the **Style** and **Size** menus.

4. Select the options you want for labeling the plot:
   - **Show Gene Names—**Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
   - **Show X Axis Label—**Displays the parameter that is graphed on the X axis.
   - **Show Y Axis Label—**Displays the parameter that is graphed on the Y axis.
   - **Show Z Axis Label**—Displays the parameter that is graphed on the Z axis.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.
5. Click **Apply**.

Changing Color Settings

Coloring in the scatter plot view is more complicated than in other views because the color of each gene can be derived from the data in any axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot lets you color genes based on a fourth experiment or condition that is not plotted on either axis.

To change color settings:

1. Open the **Display Options** window for the 3D Scatter Plot view.

   Go to "Displaying Options for the 3D Scatter Plot View" on page 8-25 for information.

   The Display Options window opens.
2. Click the **Coloring** tab.
3. In **Color data points by** menu, select the type of data that is to be used for coloring.

   For more information about the types of data that are available for coloring, see "Changing Display Options for Color Settings" on page 8-60.
4. In **Use the expression levels in the following experiment or condition** options, select the axis you want to color.

   Note that only axes which represent experiments, interpretations or conditions are available for coloring.
5. To color genes by an experiment that is not represented by either axis, do the following:
   a. Click **Other Experiment**.
   b. Select an experiment in the **Navigator**.
   c. Click **Set Experiment**.
6. Click **Apply**.

Using Error Bars

Error Bar Information displays whether the error bar is based on Standard Error, Standard Deviation or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Displaying the Average of Genes as a Centroid

When coloring by classification, GeneSpring can display the centroids for each cluster in a 2D or 3D Scatter Plot. The centroid is the average of all the genes in the cluster.

To display the average of genes as a centroid:

1. Select **View > 3D Scatter Plot**.
2. Open the Display Option for the 3D Scatter Plot.

   Go to"Displaying Options for the 3D Scatter Plot View" on page 8-25 for instructions
3. Click the **Coloring** tab.

   Go to "Changing Color Settings" on page 8-28 for more information about changing color settings.
4. In the **Color data points by** menu, select **Classification**.
5. Select the **classification** you want to use.
6. Click **Change Colors** to change the colors to display.
7. Select **OK**.
8. Select **View > Show Average of Genes**.

   All of the genes are hidden and the centroid for each cluster displays. The centroid is plotted as if it were a gene, belonging to the cluster it represents.

   Typically, only one centroid per set is shown in the Colorbar. However, if the classification used in the Colorbar is different from the classification used to split the window, there will be a centroid/average gene shown for each intersection of the Colorbar classification and the split-window classification.

   Additionally, only one centroid is displayed in the following cases:

   - The window is not split by a gene list/classification option
   - Data points are not colored by a gene list/classification option.
   - Data points are not colored by a parameter.
9. Double-click on a centroid to display the **Gene Inspector** window.

   Go to "Using the Gene Inspector" on page 7-16 for more information.

Using the Tree View

The Tree view (Figure 8-10) lets you view the results of hierarchical clustering in the form or a mock phylogenetic tree, or dendrogram. In such a tree, genes having similar expression patterns are clustered together.

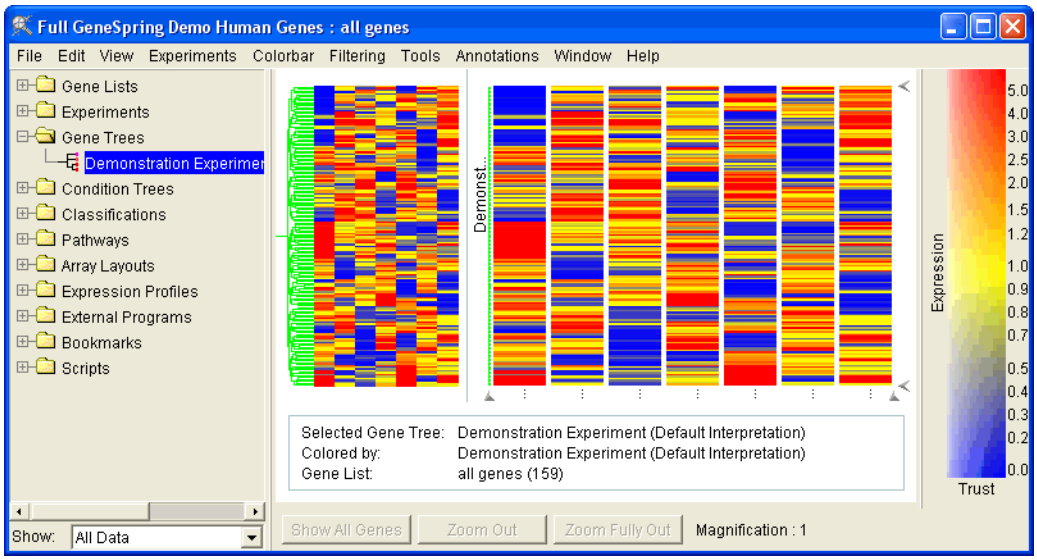


Figure 8-10 Tree View

Figure 8-10 shows a gene tree displayed in the Genome Browser. The genes are the columns of colored rectangles to the right of the tree structure, displayed in green. Similarly colored genes tend to be clustered together.

Opening the Tree View

To open the tree view:

1. In the **Navigator**, open the **Gene Trees** or the **Condition Trees** folder.
2. Select a **tree**.

   If there are no trees available for viewing, you must create one. See "Performing Gene Tree Clustering" on page 15-8.

- Select **View** > **Tree**.

Creating New Trees

To create a new tree from a node of a larger tree:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.
2. Select any node by clicking over its intersection with your cursor.
3. Right-click in the **Genome Browser** and select **Make Subtree**.

Selecting Subtrees

A single green line ending in a gene is a branch of the gene tree. Each bar crossing a set of branches forms a node of the intersecting branches. The distance from gene X to the node connecting it to gene Y indicates how closely genes X and Y are correlated. The shorter the distance, the higher the correlation.

To select a subtree:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Select any node by clicking over its intersection with your cursor.

All the genes associated with that node changes to your selected color.

Making a Gene List from a Subtree

To make a gene list from a subtree:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Select any node by clicking over its intersection with your cursor.

- Right-click in the Genome Browser and select **Make List from Subtree**.

Viewing Subtrees in the Eisen-Like Tree View

You can view subtrees in a format similar to the one generated by the visualization program by Michael Eisen. Figure 8-11 shows an Eisen-like Tree view in GeneSpring.

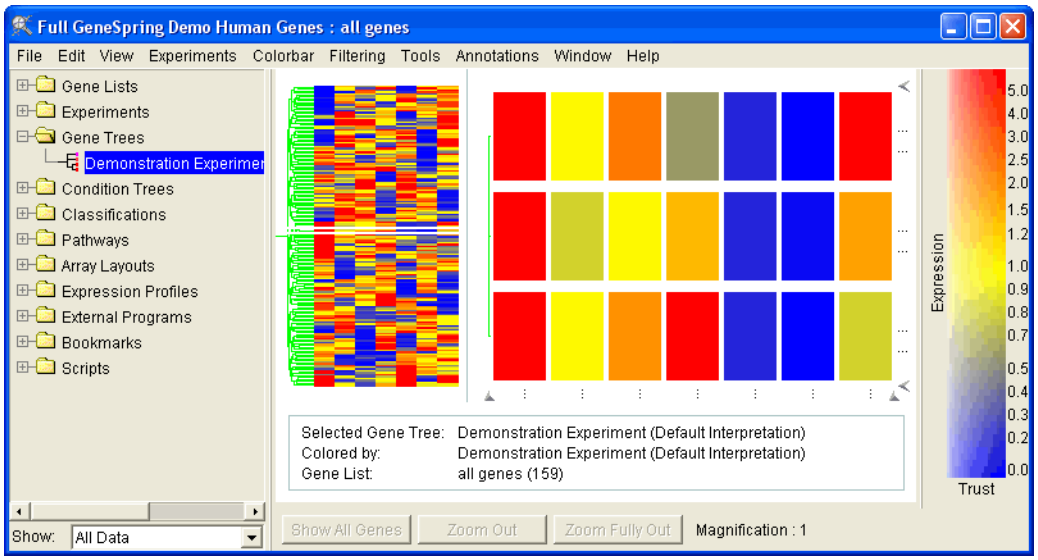


Figure 8-11 Eisen-like Tree View

Selecting Subtrees for Eisen-like Tree Views

To select a subtree for the Eisen-like Tree view:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Do one of the following:

   - Double-click on the node that defines the subtree you want.
   - Right-click on the node that defines the subtree you want and select **Display Subtree**.

3. Double-clicking a node changes the selected subtree.

   You can double-click on nodes both in the thumbnail and in the main part of the window.

Changing Thumbnails

Figure 8-12 shows three marquees in the thumbnail. The area displayed within these marquees is shown to the right of the thumbnail.

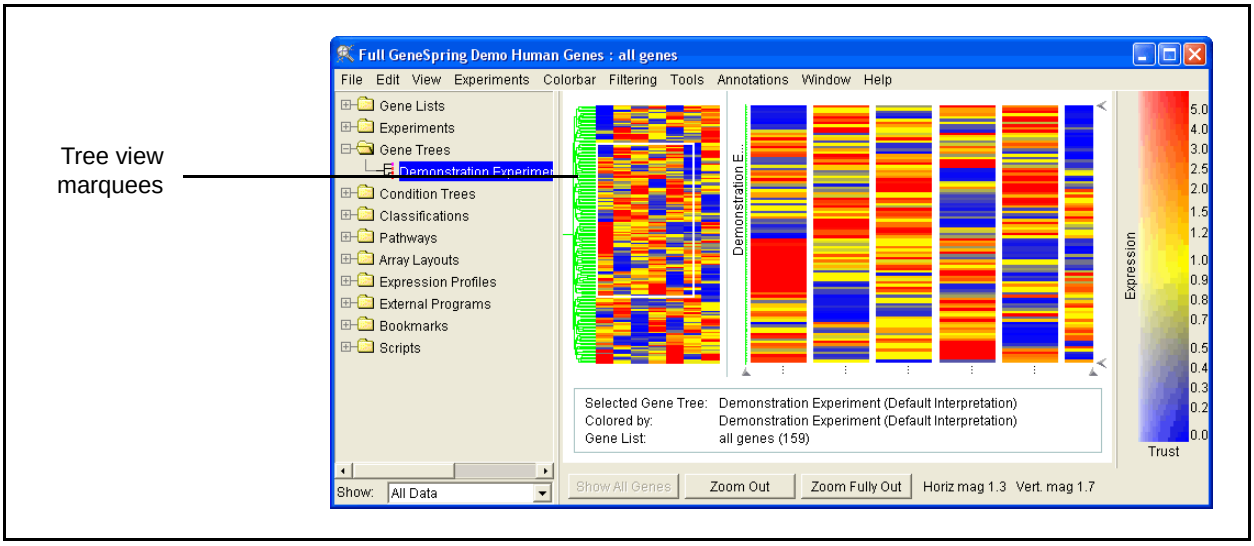


Figure 8-12 Marquees in the Tree View

To change thumbnails:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Select a subtree.

   Go to "Selecting Subtrees for Eisen-like Tree Views" on page 8-32 for instructions.

3. Use the drag arrows below and to the right of the tree.

4. To enable these drag arrows, do the following:
   a. Right-click in the browser and select **Display Options**.
   b. Click the **Features** tab.
   c. Select the **Show Drag Arrows** box.
   d. Click **OK** to return to the tree view.

Navigating Subtrees

To navigate subtrees:

1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Right-click on any node and select **Display Subtree**.
3. To view the tree immediately above the one selected, right-click anywhere and select **Display Parent of Sub-tree**.

- To return to the top and view the entire tree, right-click anywhere and select **Display Entire Tree**.

  This action returns you to the default GeneSpring Tree view.

- Double-clicking a terminal branch (a line indicating only one condition or gene) invokes either the Condition Inspector or the Gene Inspector, depending on the branch selected.

Keyboard Commands for the Right-hand or Top Tree (usually the Condition tree)

- **Alt+Left arrow**—Jump to the sibling to the left of the selected node
- **Alt+Right arrow**—Jump to the sibling to the right of the selected node
- **Alt+Up arrow**—Jump to the parent of the selected node
- **Alt+Down arrow**—Jump to the first child of the selected node (counting from left to right)

Keyboard Commands for the Left-hand Tree (usually the Gene Tree)

- **Ctrl+Left arrow**—Jump to the parent of the selected node
- **Ctrl+Right arrow**—Jump to the first child of the selected node (counting from top to bottom)
- **Ctrl+Up arrow**—Jump to the sibling directly above the selected node
- **Ctrl+Down arrow**—Jump to the sibling directly below the selected node

Magnifying Trees

Magnification in the Tree View is not quite the same as in the other views due to the need to keep the genes in the view along with the immediate tree branches. Zooming in by dragging a rectangle with the cursor usually produces a magnified view that contains more elements than were in the selected area. The amount of magnification is visible in the parameter specification area just below the Genome Browser.

Use arrow keys to pan the window while zoomed. Panning never takes you outside the bounds of the selected subtree (if any).

When a subtree is selected, clicking **Zoom Fully Out** displays the entire subtree, not the entire tree. To return to the top level, right-click anywhere and select **View Entire Tree**.

You cannot zoom in on the thumbnail in the Eisen-like Tree view.

Viewing Nodes

When you mouse over a node in the tree view, a label will appear with an indication of the distance that the node represents. The smaller the number, the more closely related are the genes in the nodes below it. This number represents the largest distance between any two genes in the node. The value is determined by the distance measure used to build the tree.

After clustering the genes according to their expression patterns, all known lists are checked against all subtrees of the new gene tree, to assign names to the tree nodes where possible. These labels are taken from the gene lists in the standard lists.

Place your cursor as close as possible to a label or intersection to view the text. When the cursor pauses over an intersection, a label appears. It disappears when the cursor is moved.

All of the branches intersecting to form a node constitute the subtree defined by that node. A label such as “ribosome [15.1]” means the subtree from that node has a lot in common with the genes in the “ribosome” list. The numbers in square brackets are a measure of statistical significance. The higher the value, the more significant the comparison is.

The comparisons between the lists and the subtrees are not looking for exact matches, but rather statistically significant overlaps, which may include subsets and supersets. When there is enough space on the window, a label, if one exists, is displayed along the top (horizontal bar) of the subtree. Otherwise, when there is space, a “...” is displayed. An “&” symbol after a list name indicates the subtree is statistically similar to more than one list, all of whom, when there is enough room, are displayed as labels along the top of the subtree.

To take a window shot that includes the label, hover your cursor over the node, take the window shot when the label appears. For most Windows applications, the cursor is not visible, just the label. For more information about window shots, see “Printing Images” on page 17-4.

Viewing Gene Names in Trees

You can magnify the tree until the names are visible along the edge of the genes.

To view gene names in trees:

1. Open the **Tree** view.

   Go to “Opening the Tree View” on page 8-30 for instructions.

2. Place your cursor anywhere over the group of genes to view the gene name.

   When the cursor pauses over a gene, a label appears. It disappears when the cursor is moved.

3. Click once and that gene becomes the selected gene.

The name of the selected gene appears in the upper right corner of the Genome Browser.

Viewing Parameters in Trees

For most experiments, each measurement was taken under certain conditions. These conditions are listed in the far right side of the tree view. If one of the parameters has been designated as a continuous parameter, it is shown directly beneath the Genome Browser.

Displaying Options for the Tree View

To display options:

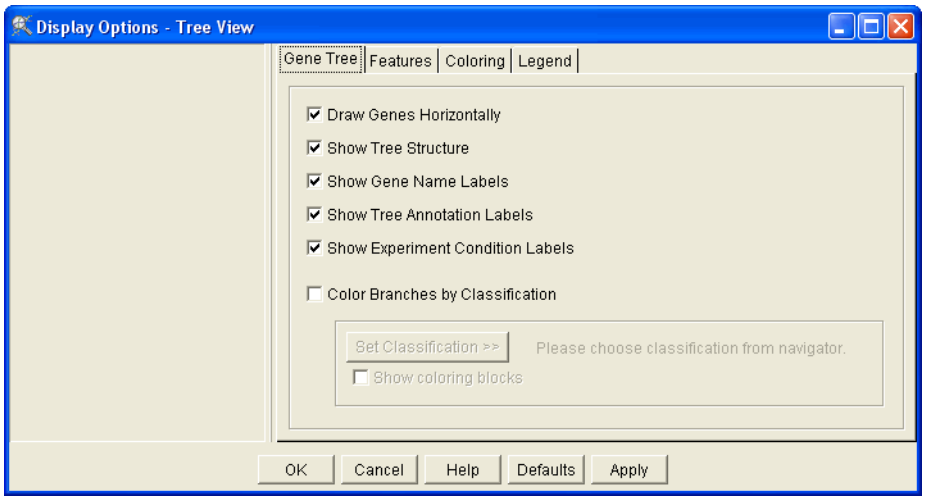
1. Open the **Tree** view.

   Go to "Opening the Tree View" on page 8-30 for instructions.

2. Do one of the following:

   - Select **View > Display Options**.
   - Right-click anywhere in the **Genome Browser**, and then select the **Display Options** command.

   The Display Options window opens.



Changing Gene Tree or Condition Tree Views

The Display Options window includes a Gene Tree or Condition Tree tab, depending on whether you have selected a gene tree, condition tree, or both. Figure 8-13 shows a Condition Tree view.

Figure 8-13 Condition Tree View

To change Gene Tree or Condition Tree views:

1. Open the **Display Options** window for the Tree view.

   Go to "Displaying Options for the Tree View" on page 8-35 for information.

   The Display Options window opens.

2. Click the **Gene Tree** or **Condition Tree** tab.

3. Select the options you want:

   - **Draw Genes Horizontally** – Orients your tree so that the genes appear as horizontal bars on the right extending from tree branches on the left.
   - **Show Tree Structure**—Specifies whether to show or hide the tree structure.
   - **Show Gene Name Labels**—If genes are displayed vertically, shows the name of each gene to its right if there is space. You must be at a very high magnification for these labels to be visible. This option is available only if a gene tree is selected.
   - **Show Tree Annotation Labels**—Displays annotations for tree nodes if they are available and space permits. This option is available only if a gene tree is selected.
   - **Show Experiment Condition Labels**—Displays experiment condition labels if they are available and space permits. This option is available only if an experiment or condition tree is selected.
   - **Color Branches by Classification**—Color the tree branches based on classification. This option is available only if a gene tree is selected.

- **Color Branches by Experiment Parameter**—Color the tree branches based on experiment parameters. This option is available only if a condition or condition tree is selected.

4. Click **Apply**.

Coloring by Classification

To color by classification:

1. Select the **Color Branches by Classification** box.
2. Select a **classification** from the Display Options mini browser.
3. Click **Set Classification**.
4. Click **Apply**.

   Unclassified genes are displayed using the background color.

Coloring by Parameter

To color by parameter:

1. Select the **Color Branches by Experiment Parameter** box.
2. Select an **experiment** from the Display Options mini browser.
3. Click **Set Experiment**.
4. Select a parameter from the menu.
5. To display a row of blocks at the bottom of the condition tree indicating their classification, select the **Show Coloring Blocks** option.
6. Click **Apply**.

Changing Features

To change features:

1. Open the **Display Options** window for the Tree view.

   Go to "Displaying Options for the Tree View" on page 8-35 for information.

   The Display Options window opens.

2. Click the **Features** tab.
3. Select the options you want:

   - **Color by all conditions**—Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the Genome Browser.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

- **Place gaps between Heatmap Tiles**—Clear this option to remove gaps between tiles.
- **Display Navigational Tree**—Specifies whether to display the navigational tree for the Eisen-like subtree view on the left or the top of the viewing area.
- **Use Custom Heatmap Borders**—lets you customize the amount of window space dedicated to tree branches and labels. When this option and the Show Drag Arrows option are selected, use the drag arrows in the Genome Browser window to make adjustments.
- **Show Drag Arrows**—Displays arrows used for changing the size of the area dedicated to tree branches. This affects both the thumbnail and the displayed subtree in the Eisen-like view.

4. Click **Apply**.

Changing Color Settings

See “Changing Display Options for Color Settings” on page 8-60.

Changing the Legend

See “Changing Display Options for the Legend” on page 8-57

Using the Ordered List View

The Ordered List view (Figure 8-14) lets you view a gene list in the order of its associated values. Values are listed in descending order. If you do not have associated values, genes are ordered according to the way they are listed in the Master Table of Genes. Vertical lines representing genes are proportional to the gene’s associated number.

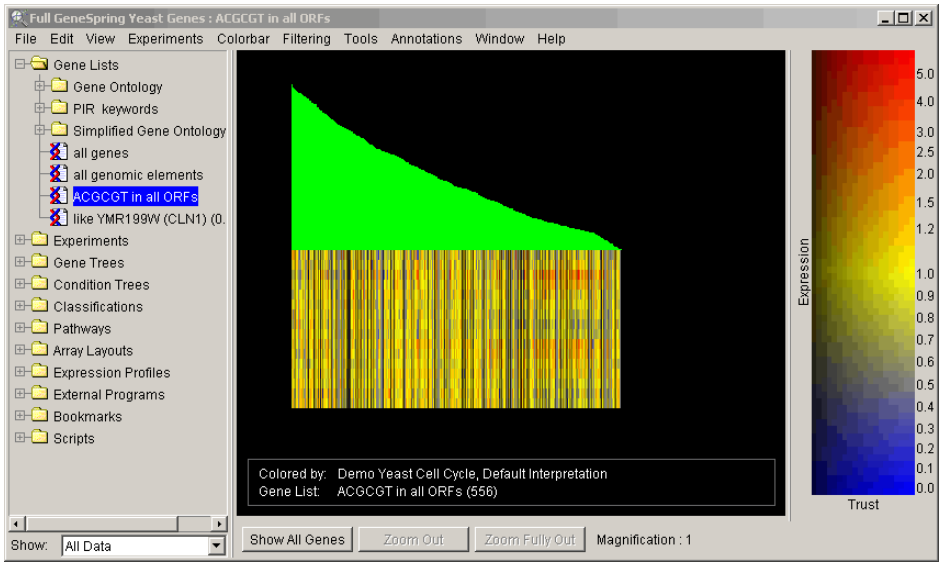


Figure 8-14 Ordered List View

Opening the Ordered List View

To open the Ordered List view:

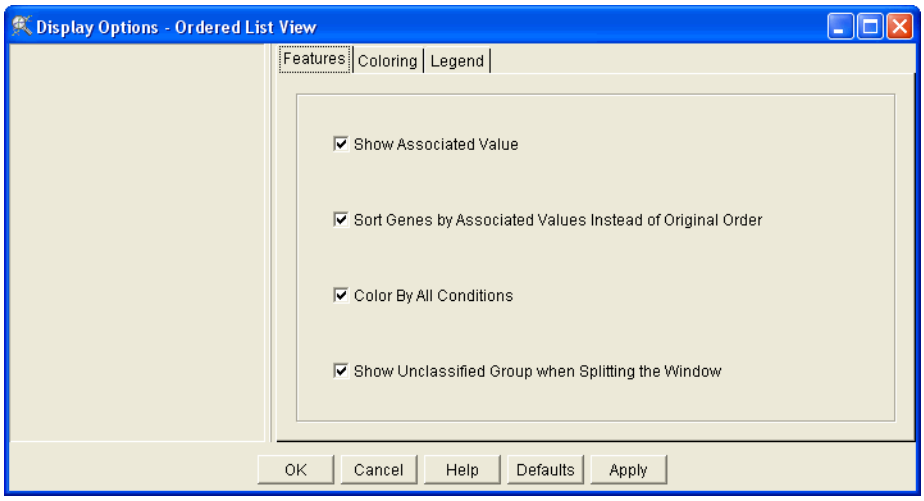
- Select **View** > **Ordered List**.

Displaying Options for the Ordered List View

To display options:

- Do one of the following:
  - Open the **Ordered List** view and then select **View** > **Display Options**.
  - Open the **Ordered List** view, right-click anywhere in the **Genome Browser**, and then select the **Display Options** command.

  The Display Options window opens.



Changing Features

To change features:

1. Open the **Display Options** window for the Ordered List view.

   Go to "Displaying Options for the Ordered List View" on page 8-39 for information.

   The Display Options window opens.
2. Click the **Features** tab.
3. Select the options you want:
   - **Show Associated Value** – When the view is zoomed, so as to enlarge the tops of the lines, selecting this options displays the numerical value associated with each line.
   - **Color by all conditions** – Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the Genome Browser.

- **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

4. Click **Apply**.

Changing Color Settings

See “Changing Display Options for Color Settings” on page 8-60.

Changing the Legend

See “Changing Display Options for the Legend” on page 8-57

Using the Array Layout View

An array contains information about the arrangement of the spots on your array. These can be used to recreate an image of your arrays to check for regional abnormalities. The Array Layout view (Figure 8-15) produces a synthetic picture of the arrays used in the current experiment. This view is useful in identifying arrays that display local shifts in intensity due to problems in probe deposition, hybridization, washing, or blocking.

To use this view you must first create an array layout file. Go to Appendix C, “Array Layout” for more information.

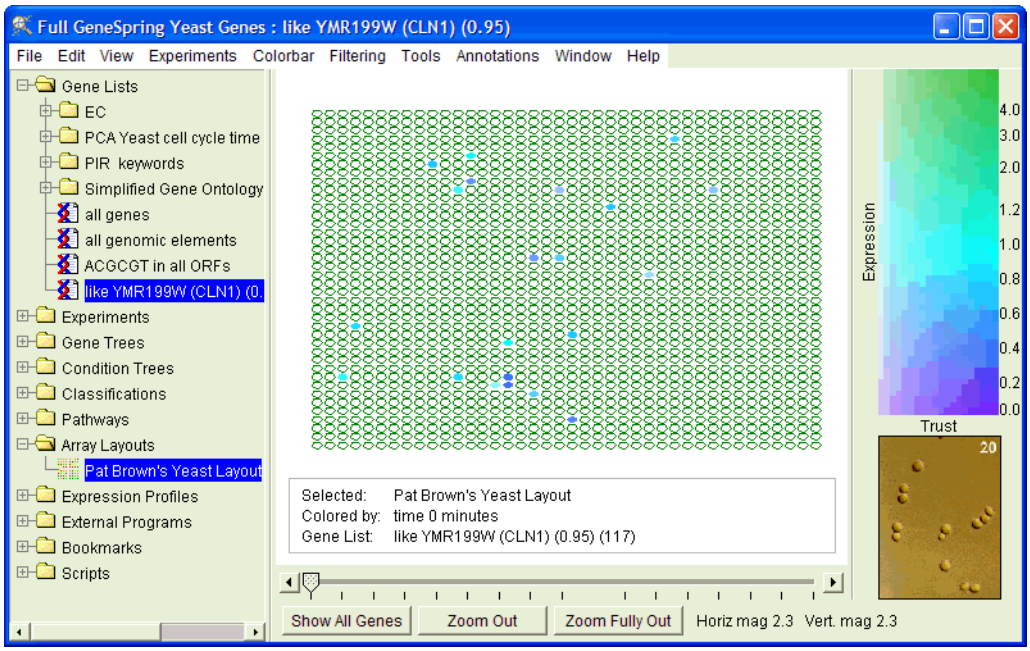


Figure 8-15 Array Layout View

In Figure 8-15, each solid circle represents an oligonucleotide on the array. If you zoom in, the gene names become visible. Circles are numbered from left to right and top to bottom. For example, a 3X3 array is:

1 2 3
4 5 6
7 8 9

Opening the Array Layout View

To open the Array Layout view:

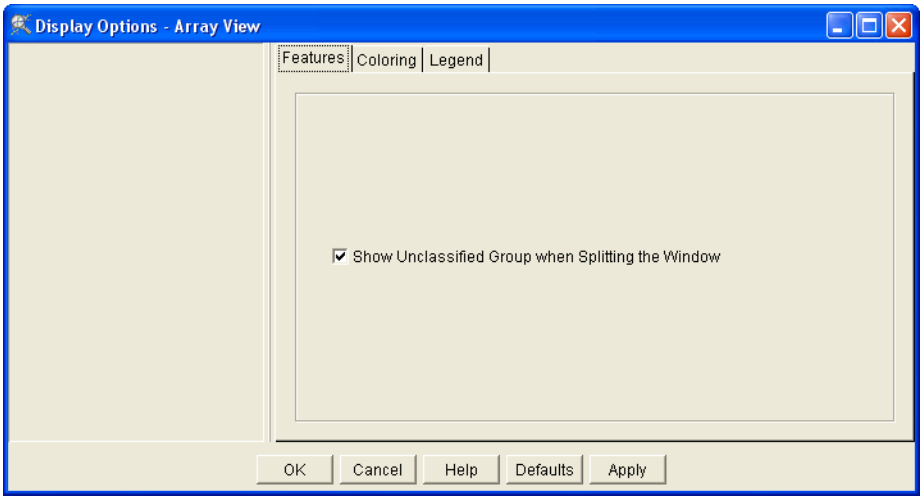
- Select **View** > **Array Layout**.

Displaying Options for the Array Layout View

To display options:

- Do one of the following:
  - Open the **Array Layout** view and then select **View** > **Display Options**.
  - Open the **Array Layout** view, right-click anywhere in the Genome Browser, and then select the **Display Options** command.

  The Display Options window opens.



Changing Features

The only feature that can be changed is the **Show unclassified Group When Splitting the Window** option within the features panel**.** When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

Changing Color Settings

See “Changing Display Options for Color Settings” on page 8-60.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Using the Pathway View

The Pathway view (Figure 8-16) lets you display and place genes on an imported gif or jpeg image. For information on downloading and importing pathways, see "Using Pathways" on page 11-10.

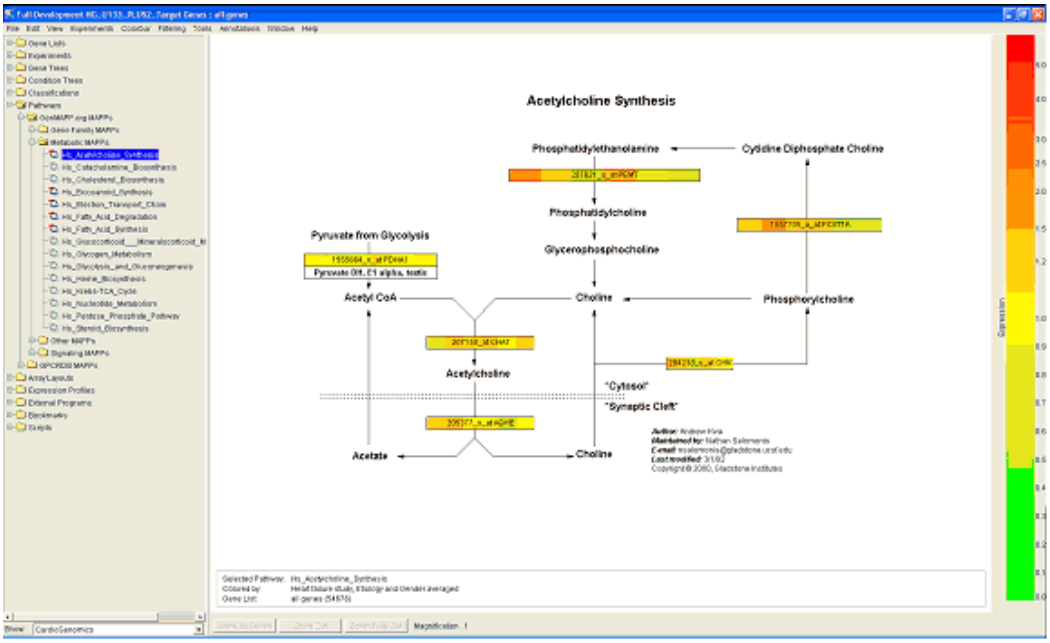


Figure 8-16 Pathway View

Opening the Pathway View

To open a pathway, you must have already created one. Go to "Using Pathways" on page 11-10 for information.

To open the Pathway view:

1. Do one of the following:
   - In the **Navigator**, open the **Pathways** folder, select the pathway you want, and then select **View > Pathway**.
   - In the **Navigator**, open the Gene Lists folder, and open the gene list you want.

   If a pathway contains a gene on a selected gene list, then the gene is colored according to its expression level in the selected experiment. See the example of the mitosis pathway in Figure 8-16.

Adding Genes to Pathways

To add a gene to a pathway:

1. Open the **Pathway** view.

   Go to "Opening the Pathway View" on page 8-42 for instructions.

2. Hold **Ctrl** and drag mouse to the area you want.
3. Type a gene name or keyword.
4. If a keyword is used, select the gene from the resulting list.

Deleting Genes from Pathways

To delete a gene from a pathway:

1. Open the **Pathway** view.

   Go to "Opening the Pathway View" on page 8-42 for instructions.

2. Right-click the gene and select **Delete Pathway Element**.

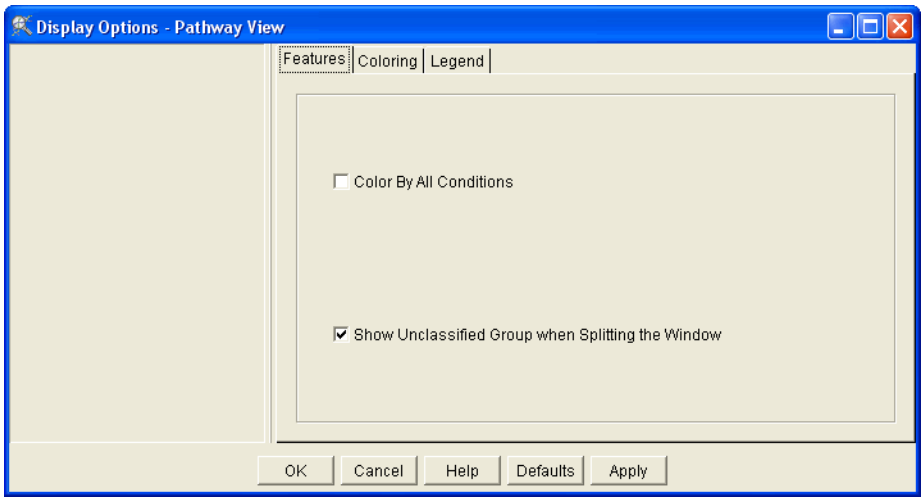
Zooming, coloration, movement and the **Find Genes Which Could Fit Here** features work in this view. **Find Genes Which Could Fit Here** suggests genes that might be appropriate in certain areas of the picture.

Displaying Options for the Pathway View

To display options:

- Do one of the following:
  - Open the **Pathway** view and then select **View > Display Options**.
  - Open the **Pathway** view, right-click anywhere in the **Genome Browser**, and then select the **Display Options** command.

  The Display Options window opens.



Changing Features

To change features:

1. Open the **Display Options** window for the Pathway view.

   The Display Options window opens.

2. Click the **Features** tab.

3. Select the options you want:

   - **Color by all conditions** –Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the Genome Browser.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

4. Click **Apply**.

Changing Color Settings

See “Changing Display Options for Color Settings” on page 8-60.

Changing the Legend

See “Changing Display Options for the Legend” on page 8-57.

Using the Compare Genes to Genes View

The Compare Genes to Genes view (Figure 8-17) lets you observe the similarity between the expression profiles of two genes in one list or in two separate lists. Genes being compared are listed along respective graph axes. The correlation between any two genes is shown by a colored square at their point of intersection. Strong correlations in expression level are shown by a higher intensity color, weak correlations by a lower intensity color.

Associated values for gene lists are shown as lines extending perpendicularly from each axis. The length of the line represents the magnitude of the associated value. You can view these associated values by zooming in on the ends of the lines.

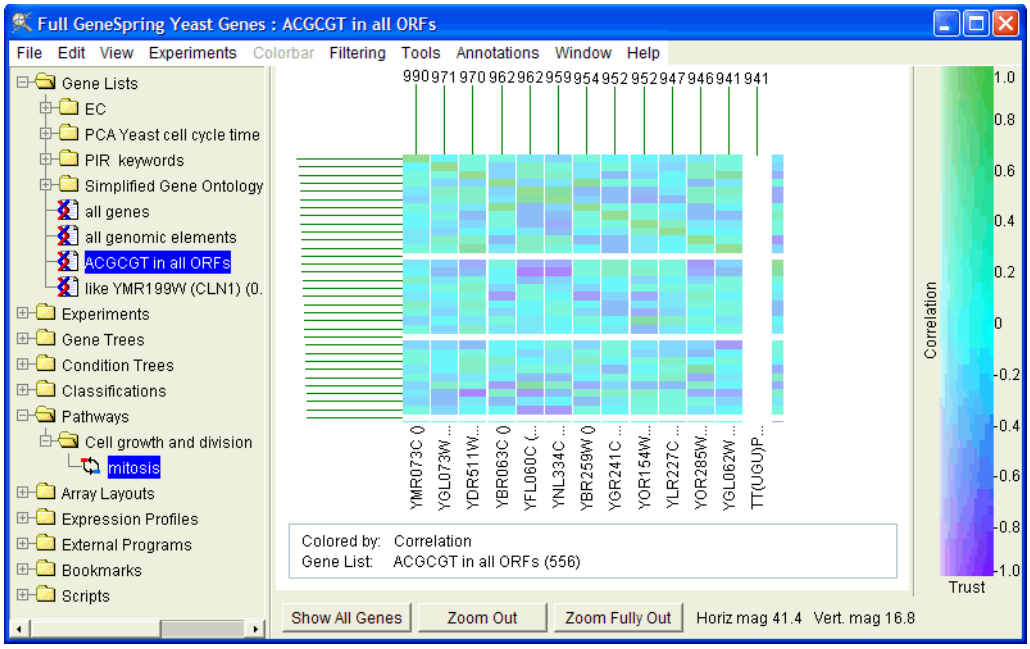


Figure 8-17 Compare Genes to Genes View

In the Compare Genes to Genes view, GeneSpring employs a Pearson correlation to measure the pair-wise similarities (see "Pearson Correlation" on page 14-16). Note that if you place the same list on both axes, a line of perfect correlation values descends diagonally across the grid. There are no display options in this view.

Opening the Compare Genes to Genes View

To open the Compare Genes to Genes view:

1. In the **Navigator**, open the **Gene List** folder.
2. Select the first gene list you want to compare.
3. Perform this step before you switch the view type, as large gene lists take a very long time to compare.
4. Select **View > Compare Genes to Genes**.

   The default display places the selected gene list on both axes.
5. (Optional) To compare a second gene list, do the following:
   a. Right-click the second gene list you want to compare.
   b. Select the **Display as Second List**.
6. To remove this second list, select **View > Remove Secondary Gene List**.

Using the Graph by Genes View

The Graph by Genes view (Figure 8-18) lets you visualize an experiment as one line, where each point on the line represents the relative expression of one gene. Genes with no data cannot be displayed in this view.

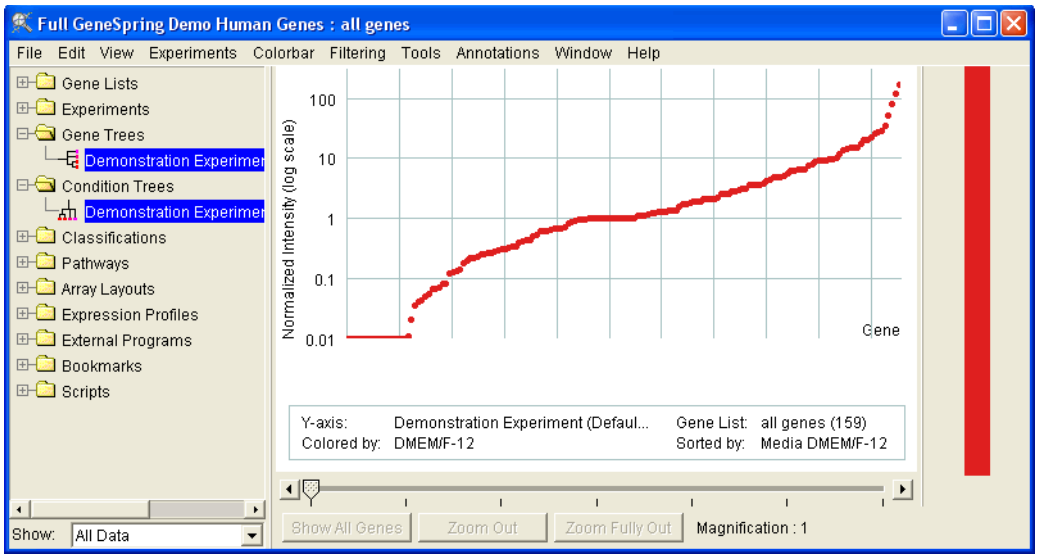


Figure 8-18 Graph by Genes View

Genes at the top of the selected gene list are displayed at the left end of the experiment line and genes at the bottom of the gene list are displayed at the right end of the experiment line. Generally, your gene lists are ordered so that the associated values appear in descending order. If you do not have associated values, your genes appears in the same order as in the master gene table.

Opening the Graph by Genes View

To open the Graph by Genes view:

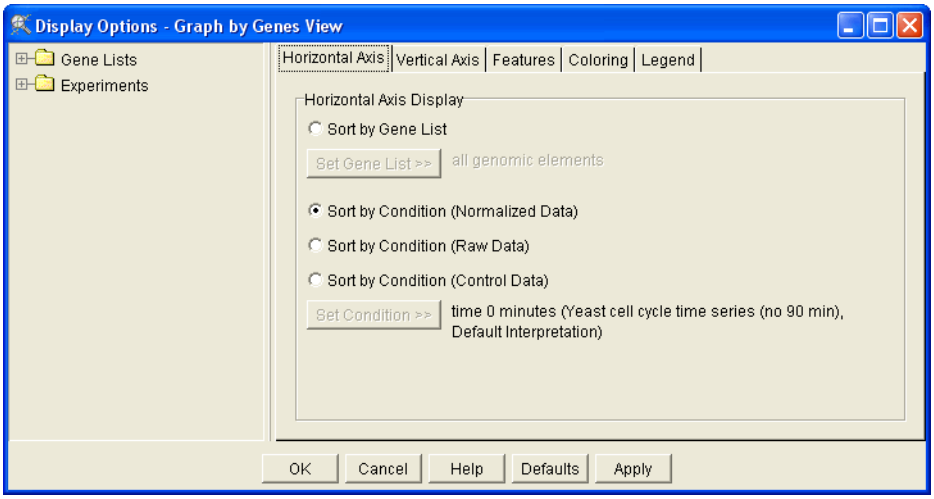
- Select **View > Graph by Genes**.

Displaying Options for the Graph by Genes View

To display options:

- Do one of the following:
  - Open the **Graph by Genes** view and then select **View > Display Options**.
  - Open the **Graph by Genes** view, right-click anywhere in the **Genome Browser**, and then select the **Display Options** command.

The Display Options window opens.



Changing the Horizontal Axis

To change the horizontal axis:

1. Open the **Display Options** window for the Graph by Genes view.

   Go to "Displaying Options for the Graph by Genes View" on page 8-46 for information.

   The Display Options window opens.

2. Click the **Horizontal Axis** tab.

3. Select the options you want.

   - **Sort by Gene List**—Sorts genes in their order in the gene list (by their associated numbers, if they exist; otherwise by their order in the Master Table of Genes).
   - **Set Gene List**—Specifies the gene list by which to sort genes. This button is only active if the **Sort by Gene List** option is selected.
   - **Sort by Condition (Normalized Data)**—Sorts genes in the order of their normalized values within the selected condition.
   - **Sort by Condition (Raw Data)**—Sorts genes in the order of their raw values within the selected condition.
   - **Sort by Condition (Control Data)**—Sorts genes in the order of their control data within the selected condition.
   - **Set Condition**—Specifies the condition by which to sort genes. This button is only active if one of the **Sort by Condition** options is selected.

4. Click **Apply**.

Changing the Vertical Axis

To change the vertical axis:

1. Open the **Display Options** window for the Graph by Genes view.

   Go to "Displaying Options for the Graph by Genes View" on page 8-46 for information.

   The Display Options window opens.

2. Click the **Vertical Axis** tab.
3. In the **Value to Graph** list, select one of the following options:
   - Normalized
   - Control
   - Raw
   - Average of Raw and Control
   - Max of Raw and Control
4. In the Graph Mode box, select one of the following options:
   - Linear
   - Logarithmic
   - Fold Change
5. To adjust the vertical axis so that all measurements are visible, check the **Scale Vertical Axis to Show all Values** box.

   The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.

6. To adjust tick spacing, do the following:
   a. Clear the **Automatic Tick Spacing on Vertical Axis** box.
   b. Enter the distance between major ticks in the **Major Tick Interval** field.
   c. Enter the number of divisions between major ticks in the **Minor Ticks per Major Tick** field.

      Note that the number of visible tick marks between major ticks is one less than the number you enter.

7. Click **Apply**.

Changing Features

To change features:

1. Open the **Display Options** window for the Graph by Genes view.

   Go to "Displaying Options for the Graph by Genes View" on page 8-46 for information.

   The Display Options window opens.

2. Click the **Features** tab.

3. Select the options you want:

   - **Plot Symbol—**Using the **Style** and **Size** menus, specify the symbol with which to display each gene. If the **Line** option is selected, individual genes cannot be selected in the Genome Browser window.
   - **Show Horizontal Axis Label—**Displays the parameter that is graphed on the horizontal axis.
   - **Show Vertical Axis Label—**Displays the parameter that is graphed on the vertical axis.
   - **Label Vertical Axis on Side—**Displays the vertical axis label vertically. If this option is unchecked, the vertical axis label sits to the right of the top of the vertical axis.
   - **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

4. Click **Apply**.

Changing Color Settings

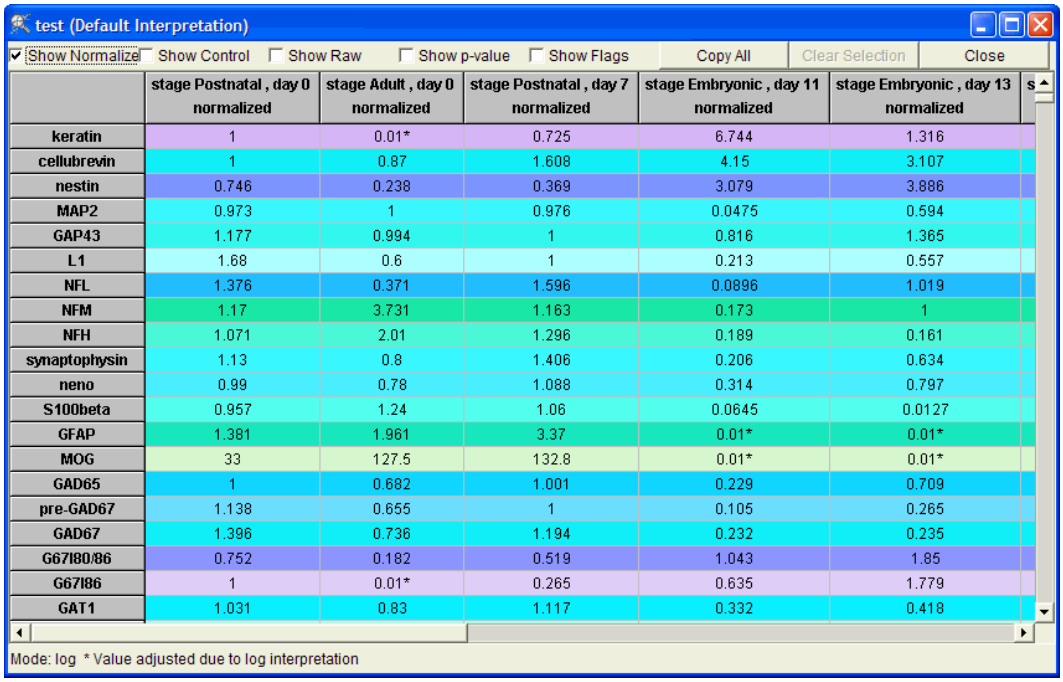
See "Changing Display Options for Color Settings" on page 8-60.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Using the Spreadsheet View

The Spreadsheet view (Figure 8-19) lets you view your data as a spreadsheet. The spreadsheet color scheme and gene list reflect what is showing in the Genome Browser at the time you activate the new window. The order of the genes is the same as in your master table of genes.



	stage Postnatal , day 0 normalized	stage Adult , day 0 normalized	stage Postnatal , day 7 normalized	stage Embryonic , day 11 normalized	stage Embryonic , day 13 normalized
keratin	1	0.01*	0.725	6.744	1.316
cellubrevin	1	0.87	1.608	4.15	3.107
nestin	0.746	0.238	0.369	3.079	3.886
MAP2	0.973	1	0.976	0.0475	0.594
GAP43	1.177	0.994	1	0.816	1.365
L1	1.68	0.6	1	0.213	0.557
NFL	1.376	0.371	1.596	0.0896	1.019
NFM	1.17	3.731	1.163	0.173	1
NFH	1.071	2.01	1.296	0.189	0.161
synaptophysin	1.13	0.8	1.406	0.206	0.634
neno	0.99	0.78	1.088	0.314	0.797
S100beta	0.957	1.24	1.06	0.0645	0.0127
GFAP	1.381	1.961	3.37	0.01*	0.01*
MOG	33	127.5	132.8	0.01*	0.01*
GAD65	1	0.682	1.001	0.229	0.709
pre-GAD67	1.138	0.655	1	0.105	0.265
GAD67	1.396	0.736	1.194	0.232	0.235
G67I80/86	0.752	0.182	0.519	1.043	1.85
G67I86	1	0.01*	0.265	0.635	1.779
GAT1	1.031	0.83	1.117	0.332	0.418

Figure 8-19 Spreadsheet View

Opening the Spreadsheet View

To open the Spreadsheet view:

- Select **View >Spreadsheet**.

Selecting Data to View

To select which data to display in the view:

1. Select **View >Spreadsheet**.
2. Select the display option you want:
   - Show Normalized
   - Show Control
   - Show Raw
   - Show p-value
   - Show Flags

Sorting Rows

To sort rows:

- Double-click the column header to sort the rows in ascending or descending order.

Copy a Row for Pasting into another Document

To copy a row:

1. Select **View >Spreadsheet**.
2. Click on the row to copy.
3. Open the document to which you want to copy the row.
4. Right-click on the row and select **Copy**.

Copying the Entire Spreadsheet

To copy the entire spreadsheet:

1. Select **View >Spreadsheet**.
2. Click **Copy All**.

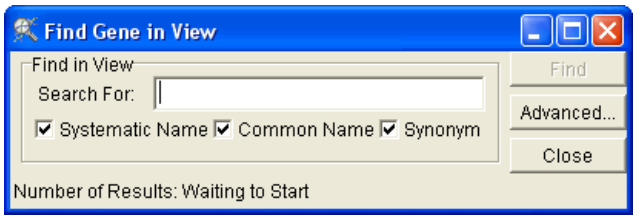
Note: If you have any rows selected, you must first click **Clear Selection**.

Finding a Gene

To find a gene:

1. Select **View >Spreadsheet**.
2. Do one of the following:
   - Click **Find**.
   - Type **Ctrl+F**.

   The Find Gene in View window opens.



3. Enter the gene name.
4. Select the annotations you want.
5. Click **OK**.

Inspecting Genes

To bring up the Gene Inspector for the gene you found, type **Ctrl+I**.

Using the Condition Scatter Plot

The Condition Scatter Plot view (Figure 8-20) displays a fundamentally different type of information than any other view with the possible exception of the condition tree. Unlike other GeneSpring views, each colored point (dot, circle, square, etc.) represents a condition, not a gene.

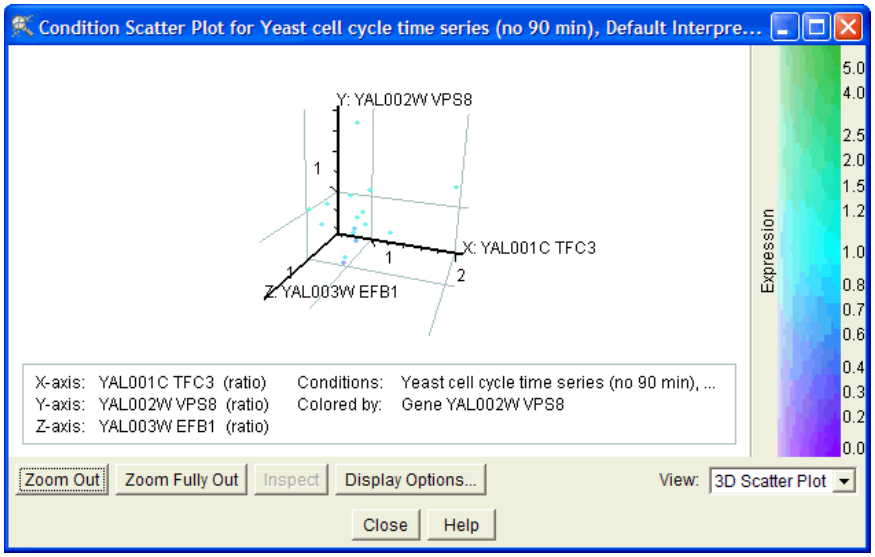


Figure 8-20 Condition Scatter Plot View

This view is the most common way to visualize the results of principal components analysis performed on conditions. It is also useful for presenting complex multidimensional data in the context of conditions.

In Figure 8-20, each dot represents a condition. When this window is opened from the main GeneSpring window, the first three parameters (if available) are selected for the axes. If only two parameters are available, the plot is displayed in 2D format. If there are fewer than two parameters, a 3D plot is displayed using the first three genes from the selected gene list.

A 3D Condition Scatter Plot can be configured to display a principal component score on one axis, a parameter value an a second axis, and the normalized expression level of a given gene on the third axis. A simpler possibility is to plot the expression values for two genes on two axes. Such a plot is useful for demonstrating whether the expression pattern of the genes is correlated or anti-correlated.

Opening the Condition Scatter Plot View

Unlike most views, the condition scatter plot is displayed in a separate window, which appears when the option is selected.

Note: This window also appears when you run a PCA on Conditions analysis. For details, see "Running a PCA on Conditions" on page 15-23.

To open the Condition Scatter Plot view:

- Select **View > Condition Scatter Plot**.

To open a 2D view of this plot:

- Select **2D Scatter Plot** from the **View** menu in the lower right portion of the window.

  Note: You cannot select the experiment to be displayed from within this view. To change the experiment being viewed, exit this window, select the experiment you want, and choose **View > Condition Scatter Plot** in the main GeneSpring window.

Rotating the Condition Scatter Plot View

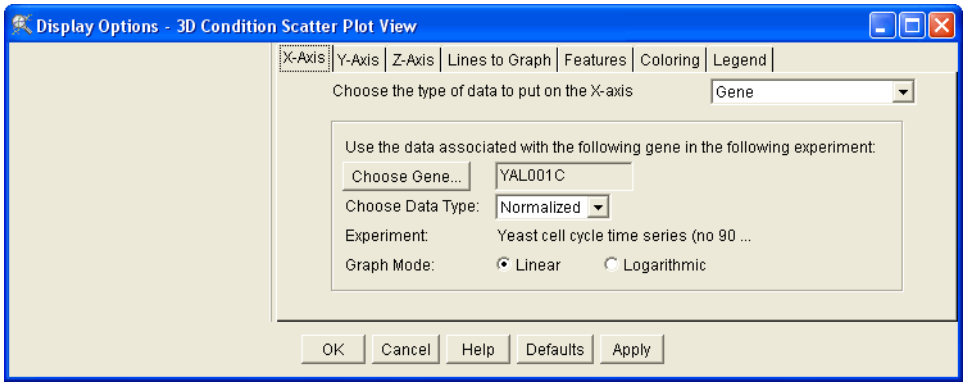
Pressing the X, Y, or Z keys rotates the graph on the specified axis. Hold down the Shift key to speed this rotation. Hold down the Alt key to reverse the direction of rotation.

Displaying Options for the Condition Scatter Plot View

To display options:

- Do one of the following:
  - Open the **Condition Scatter Plot** view and then select **View > Display Options**.
  - Open the **Condition Scatter Plot** view, right-click anywhere in the **Genome Browser**, and then select the **Display Options** command.

  The Display Options window opens.



Changing the X, Y, and Z Axes

The most critical option to set is the type of data that is displayed on the three axes.

To change the axes:

1. Open the **Display Options** window for the Condition Scatter Plot view.

   Go to "Displaying Options for the Condition Scatter Plot View" on page 8-53 for information.

   The Display Options window opens.

2. Click the **X Axis**, **Y Axis**, or **Z Axis** tab.
3. Specify the type of data to display on the selected axis from the menu. The options are:
   - Gene
   - Expression Profile
   - Experimental Parameter
4. Select the gene from which to use data in the plot:
5. Click **Choose Gene.**

   The Find Target Gene window opens.

6. Select the search criteria you want and click **Find**.

   Note: You can select genes only from the currently active experiment. To work with data from a different experiment, you must exit this screen and select that experiment in the main GeneSpring window before re-opening the Condition Scatter Plot window.

7. In the **Choose data type** option, select the option you want. The options are:
   - Control
   - Raw
   - Normalized
8. In the **Graph Mode** option, choose the option for the specified axis. The options are:
   - Linear
   - Logarithmic
   - Fold Change

   Note: The fold change option is only available if you are looking at normalized data from an interpretation or a condition.

9. Click **Apply**.

Changing Lines to Graph

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines.

These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the Genome Browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit.

To change lines:

1. Open the **Display Options** window for the Condition Scatter Plot view.

   Go to "Displaying Options for the Condition Scatter Plot View" on page 8-53 for information.

   The Display Options window opens.

2. Click the **Lines to Graph** tab.
3. To see a grid inside the plot area, select the **Horizontal/Vertical Grid Lines** check boxes.You can have lines drawn at the major and minor tick intervals of each axis.

   The color of these grid lines is represented in the Grid Color box at the bottom of the window.

4. To modify the grid color, click **Change**.
5. Click **Apply**.

Changing Features

The 3D Scatter Plot view also lets you change the appearance of data points and data labels.

To change features:

1. Open the **Display Options** window for the Condition Scatter Plot view.

   Go to "Displaying Options for the Condition Scatter Plot View" on page 8-53 for information.

   The Display Options window opens.

2. Click the **Features** tab.
3. To modify the size and shape of the points, select the option you want in the **Style** and **Size** menus.
4. Select the options you want for labeling the plot:

   - **Show Gene Names—**Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
   - **Show X Axis Label—**Displays the parameter that is graphed on the X axis.
   - **Show Y Axis Label—**Displays the parameter that is graphed on the Y axis.
   - **Show Z Axis Label**—Displays the parameter that is graphed on the Z axis.

- **Show unclassified Group When Splitting the Window—**When the window is split, this option displays the genes that were not put into any classification into their own section of the Genome Browser.

5. Click **Apply**.

Changing Color Settings

Coloring in the scatter plot view is more complicated than in other views because the color of each gene can be derived from the data in any axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot lets you color genes based on a fourth experiment or condition that is not plotted on either axis.

To change color settings:

1. Open the **Display Options** window for the Condition Scatter Plot view.

   Go to "Displaying Options for the Condition Scatter Plot View" on page 8-53 for information.

   The Display Options window opens.

2. Click the **Coloring** tab.
3. In the **Color Conditions by** menu, select the type of data that is to be used for coloring.

   For more information about the types of data that are available for coloring, see "Changing Display Options for Color Settings" on page 8-60.

4. To modify the grid color, click **Change**.
5. Click **Apply**.

Using Error Bars

Error Bar Information displays whether the error bar is based on Standard Error, Standard Deviation, or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information.

Changing the Legend

See "Changing Display Options for the Legend" on page 8-57.

Changing Common Display Options

This section explains how to change display options that are common to most views in GeneSpring.

Opening the Display Options Window

To open the Display Options window:

1. In the **View** menu, select the view you want.
2. Select **View > Display Options**.

   The Display Options window opens.

Changing Display Options for the Legend

You can specify what information to display in most views using the Legend tab (Figure 8-21) on the Display Options window. The options that are available depend on whether they are applicable to the current view.

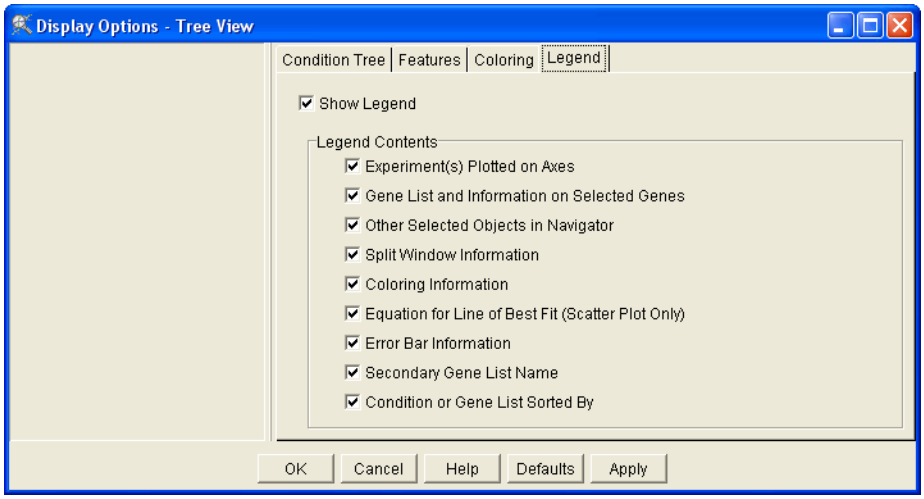


Figure 8-21 Display Options: Legend Tab

Legend Options

This section describes the different options in the Legend tab.

Selected Object in Navigator

Selected Object in Navigator displays the following:

- In Tree view, the name of the selected gene tree and condition tree.
- In Array Layout view, the name of the array layout.
- In Pathway view, the name of the pathway.

Experiment(s) Plotted on Axis

Experiment(s) Plotted on Axis displays the following:

- The name of the experiment and interpretation being displayed on the Y Axis. This is available in the following views: Graph, Graph by Genes, Bar Graph, Scatter Plot, and 3D Scatter Plot.
- The name of the experiment and interpretation (or gene list and type of associated values) being displayed on the X Axis. This is available in the Scatter Plot and 3D Scatter Plot views.
- The name of the experiment and interpretation (or gene list and the type of associated values) being displayed on the Z Axis. This is available in the Scatter Plot and 3D Scatter Plot views.

Split Window Information

Split Window Information displays the name of the Classification or Gene List folder used to split the window.

Coloring Information

Coloring Information displays different information depending on the coloring scheme selected:

- **Color by Expression**—Shows the name of the experiment, interpretation, and condition used for coloring. For experiments with continuous numeric parameters, the "condition" may actually be an interpolation between two measured conditions. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.
- **Color by Significance**—Shows the name of the experiment, interpretation, and condition used for coloring. For experiments with continuous numeric parameters, the "condition" may actually be an interpolation between two measured conditions. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.
- **Venn Diagram**—Displays "Venn Diagram"
- **Color by Parameter**—If the experiment has parameters designated as color codes, displays the name of the experiment, interpretation, and parameter(s) used for coloring. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.
- **Color by Classification**—Shows the name of the Classification or Gene List Folder used for coloring.

Equation for Line of Best Fit

Equation for Line of Best Fit applied to the Scatter Plot view only. If Line of Best Fit is selected, displays the equation for the line of best fit (data-dependent).

Error Bar Information

Error Bar Information displays whether the error bar is based on Standard Error, Standard Deviation or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information. This is available in the following views: Graph, Graph by Genes, Bar Graph, Scatter Plot, and 3D Scatter Plot.

Gene List and Information on Selected Genes

Gene List and Information on Selected Genes displays the name of the selected gene list, the number of genes in the gene list, and the name of the selected gene (if only one is selected) or the number of selected genes (if multiple are selected).

Secondary Gene List Name

If a secondary gene list is being displayed, Secondary Gene List Name displays the name of the secondary gene list and the number of genes in this list

Condition or Gene List Sorted By

In the Graph by Genes View, Condition or Gene List Sorted By displays the name of the gene list or the name of the condition, experiment and interpretation used to sort the genes on the X Axis.

Changing the Legend

To change the legend:

1. Open the Display Options window.
2. Click the **Legend** tab.
3. Do one of the following:
   - Check the **Show Legend** box to display the information you want.
   - Clear the check box, to hide the information.
4. Click **Apply**.

   Your changes are applied to the display in the main GeneSpring window.

Changing Display Options for Color Settings

The Coloring tab (Figure 8-22) of the Display Options window lets you change color settings for selected objects. The options that are available depend on whether they are applicable to the current view. This section covers the following topics:

- Changing Colors by Expression
- Coloring by Significance
- Coloring by Venn Diagram
- Coloring by Parameter
- Removing Color
- Applying a Solid Color
- Coloring by Classification
- Coloring by Split Window and Classification
- Coloring by Secondary Experiment
- Changing the Default Colors

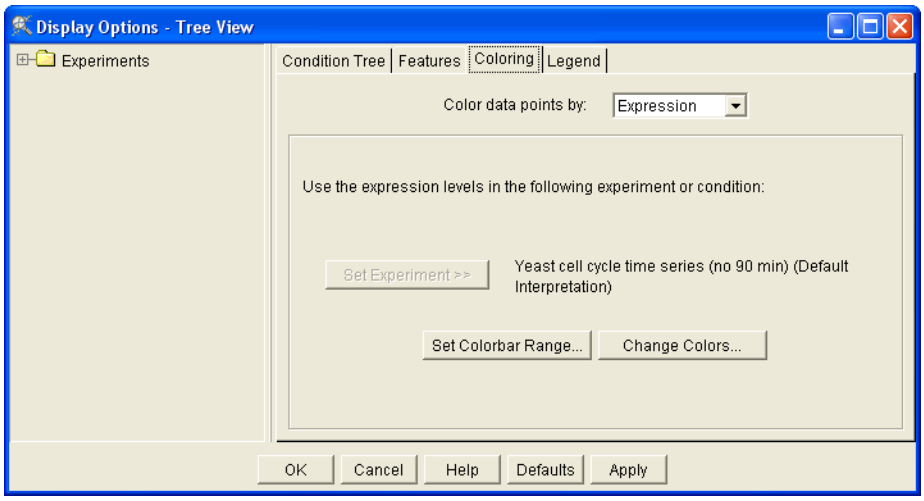


Figure 8-22 Display Options: Coloring Tab

Changing Colors by Expression

The Change Colors option colors genes according to their normalized expression values and trustworthiness. Figure 8-23 shows these elements in the Genome Browser.

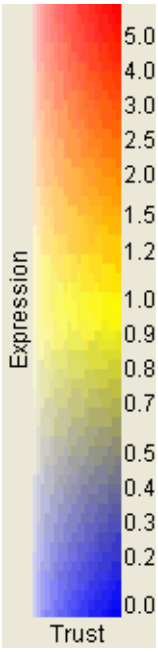


Figure 8-23 Expression and Trust

Expression

The vertical axis of the Colorbar represents expression levels on a continuous scale. Using the default colors, red indicates overexpression, yellow indicates average expression, and blue indicates underexpression. Genes are colored by their expression level in the selected condition as indicated by the condition line. If you have specified the parameter on the horizontal axis to be continuous, expression levels in between conditions are interpolated.

Trust

The horizontal axis of the Colorbar indicates the degree to which you can trust your data, where dark or unsaturated colors represent low trust, and bright, saturated colors represent high trust. GeneSpring uses the following guidelines to automatically create trust values:

- In **two-color experiments**, the trust value is usually the control channel (typically Cy5), unless you do a per chip normalization in which case it is:

  (the control channel) x (the median of the control channel) x
  (the median of the signal channel)

- For **Affymetrix and other one-color experiments**, the trust value is constructed based on the normalizations you have chosen. If you accept the default normalizations for Affymetrix data (use distribution of all genes using the 50th percentile and normalize to the median for each gene), then trust is:

  (the median value of the chip) x (the median value of the gene)

- If you choose to use distribution of all genes using the 50th percentile and normalize to sample(s), trust is calculated as follows:

(the median value of the chip)
x
(the average of the gene's measurement in control samples)

Coloring Genes by Expression

To color genes by expression:

- Do one of the following:
  - Select **View > Display Options** > **Coloring**, and then select **Expression**.
  - Select **Colorbar > Color by Expression**.

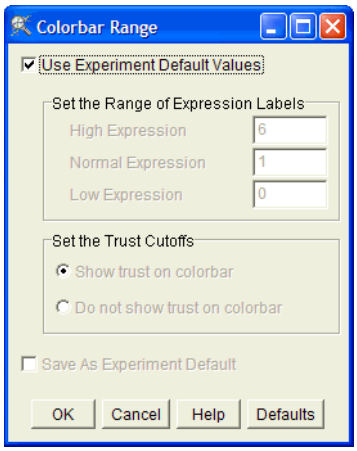
Setting Trust Interpretation

To set the trust interpretation:

1. In the **Genome Browser**, right-click the Colorbar.
2. Click **Set Coloring**.

   The Display Options window opens, with the Coloring tab pre-selected.
3. Click **Set Colorbar Range**.

   This button is active only when coloring by expression. The Colorbar Range dialog appears.



4. To enable custom settings, clear the **Use Experiment Default Values** box.
5. Enter values for High Control Strength, Medium Control Strength, and Low Control Strength.
6. (Optional) By default, trust is shown on the Colorbar. To disable the default, select the **Do not show trust on Colorbar** option.
7. To save these settings as the default for this experiment, select the **Save As Experiment Default** box.

If you leave this box unchecked, your changes will affect the display options only in your current session.

8. Click **OK**.

Changing the Colorbar Range

To change the Colorbar range:

1. In the **Genome Browser**, right-click the Colorbar, and then select **Set Coloring**.
2. Select **Color by Expression** from the menu.
3. Reset the values determining the intensity of the colors used by the Genome Browser.

   There are six categories you can change:

   - **High Expression—**High expression refers to the normalized expression of your genes, it is the vertical axis of the color bar. The default for this is 6.0.
   - **Normal Expression—**For most normalization procedures the data are normalized to 1.0. The default for this is 1.0.
   - **Low Expression—**For most normalization procedures the data do not have negative numbers. The default for this is 0.0.

   For example, you could change the usual range of an experiment to high = 10, normal = 5 and low = -2 resulting in a very different color scheme once you click the **OK** button.

   There is no **Edit > Undo** (**Ctrl+Z**) function for this type of change. To return to your previous coloration scheme, you must re-open the Experiment Data Range pop-up window and enter your old values.

   For more details on trust, see “Trust” on page 8-61. For more details on normalization, see Chapter 10,“Normalizing DataŽ’.

4. Click **OK**.

Coloring by Significance

Data are colored based on how far the gene is over- or underexpressed (relative to a normalized expression level of 1), in terms of the standard error of the measurement. The standard Colorbar is replaced with a Colorbar ranging from +3σ to -3σ.

The standard error model is based on the Cross-Gene Error Model, if the Cross-Gene Error Model is turned on. (For more information about the Cross-Gene Error Model, see “Using Cross-Gene Error Models” on page 5-73.) Otherwise the standard error is based on the standard deviation of the replicate data for a particular gene and condition (for information about the calculation of this error, see “Using the Gene Inspector” on page 7-16).

To color genes by significance:

- Do one of the following:
  - Select **View > Display Options** > **Coloring**, and then select **Color by Significance**.
  - Select **Colorbar > Color by Significance**.

Coloring by Venn Diagram

The Color by Venn Diagram option colors genes based on their membership in one or more gene lists in a Venn diagram.

To assign a gene list to the Venn diagram:

1. Do one of the following:
   - Select **View > Display Options** > **Coloring**, and then select **Color by Venn Diagram**.
   - Select **Colorbar > Color by Venn Diagram**.
2. Drag the list from the Navigator to the appropriate section of the venn diagram at the right side of the window.

You can also assign the circles in a Venn diagram by right-clicking on a gene list and selecting the **Venn Diagram** option. For more information about creating Venn diagrams and using them for analysis, see “Making Gene Lists from Venn Diagrams” on page 12-18.

Coloring by Parameter

This option colors genes based on the value of parameters. This coloring scheme is best suited for use with Graph view and Bar Graph view (Figure 8-24) when different conditions are indicated with discrete symbols.

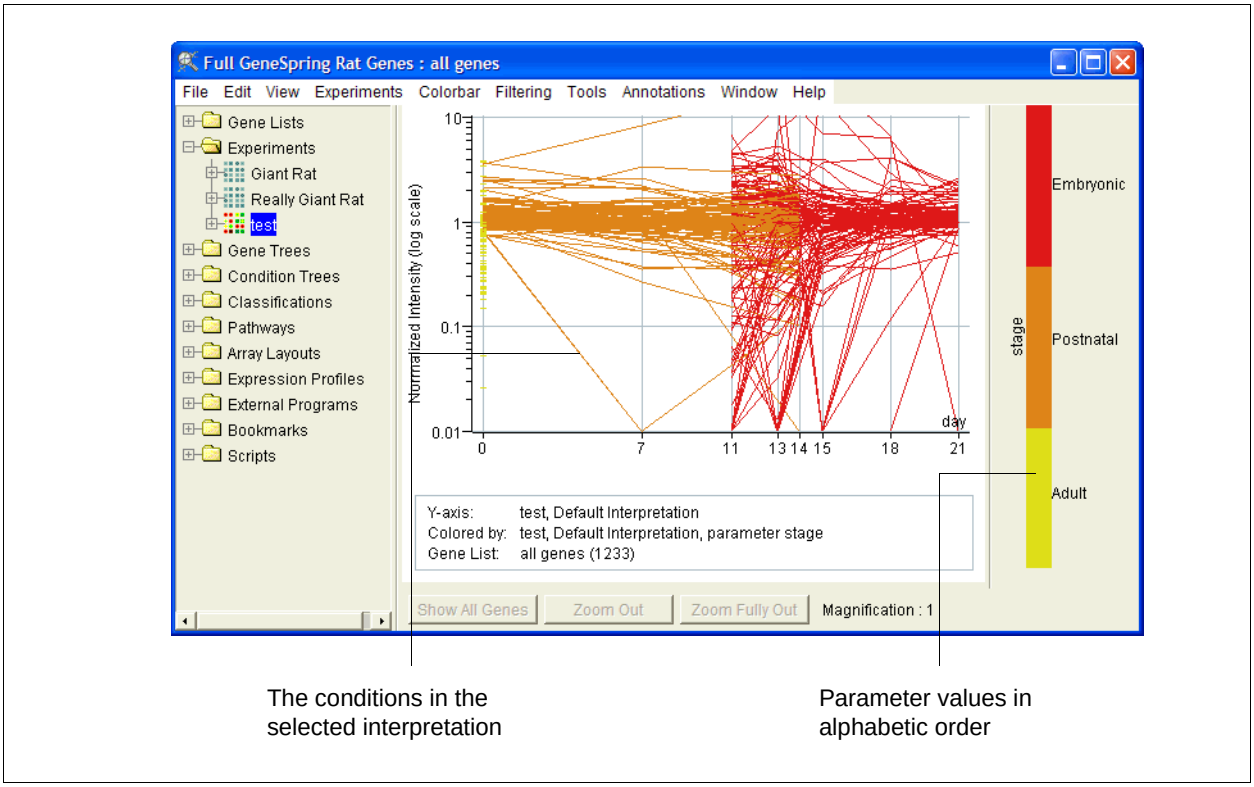


Figure 8-24 Graph View: Coloring by Parameter

To color by parameter:

1. Select **Experiments > Change Experiment Interpretation**.
2. Choose the parameter(s) to color.
3. Click **Color Code** for that parameter.
4. Click **Save** to create a new interpretation.
5. Select **Colorbar > Color by Parameter**.

To color by parameter using the Display Options window:

1. Select **View > Display Options** and click the **Coloring** tab.
2. Select a parameter from the **Navigator** on the left side of the Display Options window.
3. Click **Set Experiment**.
4. Click **OK**.

Removing Color

This option allows you to view genes with no coloration, showing all genes in gray. To implement this option, select **Colorbar > No Color**.

Applying a Solid Color

You can also select a single color in which to display genes by selecting the **Solid Color** option from the menu on the Coloring tab of the Display Options window.

Coloring by Classification

This coloring scheme lets you color-code the genes by some previously defined knowledge about them. You can use a folder of lists to color by classification or a classification method such as k-means or SOM.

To color by a previously saved classification:

1. In the **Navigator**, open the **Classifications** folder.
2. Select a classification by right-clicking the name.
3. Select **Use Coloring** from the menu.

   GeneSpring automatically updates to reflect the new coloring scheme. The Colorbar shows the names of the sets present in the chosen classification.

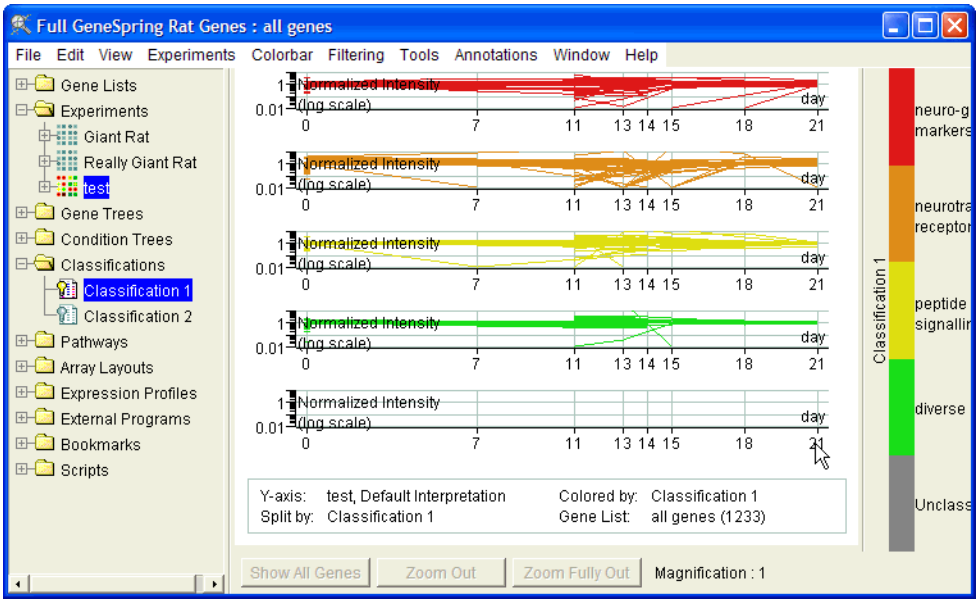


Figure 8-25 Example Split Window Colored by Classification

To color by classification using the Display Options window:

1. Select **View > Display Options** and click the **Coloring** tab.
2. Select a classification from the **Navigator** on the left side of the **Display Options** window.
3. Click **Set Experiment**.
4. Click **OK**.

Coloring by Split Window and Classification

You can use the Split Window feature with the Color by Classification scheme.

1. In the **Navigator**, open the **Gene List** folder.
2. Select a gene list to view.
3. Right-click a folder or a previously saved classification and select **Use as Classification**.
4. Right-click on the folder again and select **Split Window > Both**.

Coloring by Secondary Experiment

The Graph and Scatter Plot views lend themselves to being colored in many different ways because the display presents expression levels of the genes through the entire experiment. These are the only views in which you may choose to color the genes by a secondary experiment. This means the color of each gene line graphed correlates to the expression level of that gene in a different experiment, at the point in the second experiment marked by the secondary scroll bar.

To color by a secondary experiment:

1. In the **Navigator**, open the **Experiments** folder.
2. Position your cursor over an experiment (not the one currently displayed) you want to use for coloration.
3. Right-click and select **Set Secondary Experiment** from the menu.

The coloring scheme of the Genome Browser is shown in the Colorbar on the right. There are two versions of the animation controls in the Experiment Specification Area.

Changing the Default Colors

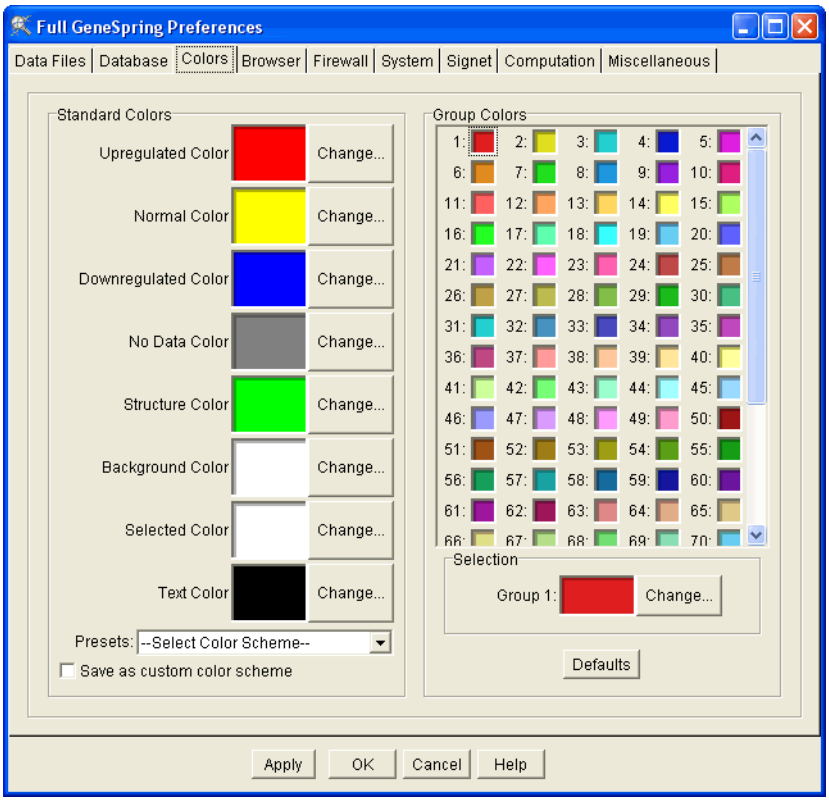
You can change the colors used to display the genes. This does not affect interpretation of your data, but it can help you to make genes more visible on-window or make it easier to print window shots.

To change the default colors:

1. Select **Edit > Preferences**.

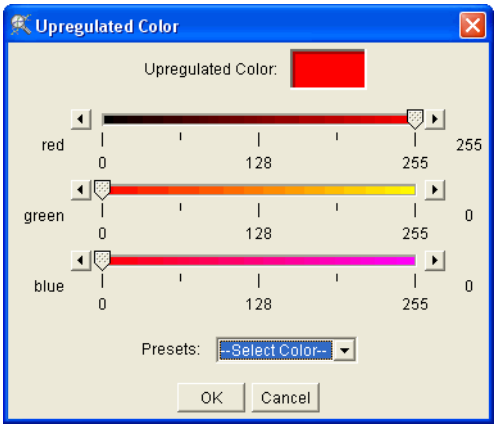
   The Preferences window opens.
2. Click the **Colors** tab.
3. Click the **Change Colors** button.

   The Preferences window opens.



4. Locate the object whose color you want to change and click the **Change** button.

   The Change Colors window opens.



5. Adjust the sliders until the color you want is displayed in the preview window at the top of the Structure Color window.

6. Click **OK**.

For more details about the other options in the Preferences window, see "Setting Preferences" on page 3-15.

9

Filtering Data

This chapter explains how to use the basic and advanced filtering tools in GeneSpring. It covers the following topics:

- Filtering
- Filtering on Gene Lists
- Using Advanced Filters
- Filtering Data Objects Assigned to Projects

Filtering

Using GeneSpring's sophisticated filtering tools, you can identify genes that are affected by novel drug treatments or experimental conditions. A variety of intuitive visual interfaces allow even novice users to select genes with specific expression patterns.

GeneSpring offers visually-intuitive filtering tools for both entry-level and advanced users. All visual filtering windows generate graphs of results in real-time. These filters allow researchers to exclude particular conditions, set minimum and maximum values, and choose specific gene lists to filter.

GeneSpring also has an advanced filtering window designed for power users. The advanced filtering window allows you to create complex Boolean expressions to identify genes with a highly-specific expression pattern.

Once created, filters can easily be saved to standardize critical laboratory procedures, or can be shared with other researchers using Signet.

Filtering on Gene Lists

Gene filtering is a simple, but effective way to sort through the large amounts of expression data. Filtering enables you to evaluate the quality of sample before performing data analysis or identify interesting genes for further study after analysis. This section includes the following topics:

- Gene Filters
- Filtering Menu
- Filter Window
- Data Types for Restrictions

- Filtering on Expression Level
- Filtering on Fold Change
- Filtering on Error
- Filtering on Confidence
- Filtering on Parameters
- Filtering on Flags
- Filtering on Gene List Numbers

Gene Filters

As a highly versatile data mining tool suitable for both quality assessment of expression measurements and analysis, filters let you identify:

- Genes that fall below a given intensity value threshold
- Data that exceeds recommended signal-to-noise or signal-to-background measurements
- Outliers that fall outside the range of standard deviations from the mean
- Random quantification errors
- Genes that do not show any expression changes during the experiment
- Interesting genes suitable for additional analyses

Filters can be applied to one or multiple data objects. Genes can be filtered on specific expression criteria and the genes that pass a filter are made into a gene list. The filters are based on any data associated with the genes including raw or normalized intensity values, fold change comparisons, flag values, statistical information, or raw data from the scanning software.

Filtering Menu

The Filtering menu lets you apply a series of restrictions or filters to a gene list. These restrictions can apply to an entire experiment or interpretation, or to a single condition or sample. The filters include factors such as quality control, control strength, expression level constraints, sample to sample fold comparison, statistical group comparisons, and associated numbers restrictions. All restrictions applied to create a new list are saved in the notes.

The ability to restrict a gene list based on the behavior of its genes in experiments or in individual samples is an important quality control tool. You may want to remove genes with low precision, large error values, those that do not vary significantly across multiple samples, or those with expression levels that are too close to the background. Filtering genes also lets you search for genes that are differentially expressed over two or more conditions.

The Filtering menu provides the following gene lists filters:

- Filter on Expression Level
- Filter on Fold Change
- Filter on Error
- Filter on Confidence
- Filter on Parameter
- Filter on Flags
- Filter on Data File
- Filter on Arbitrary File
- Filter on Gene List Numbers
- Advanced Filtering

Filter Window

The Filter window (Figure 9-1) contains elements that are similar for all gene list filters. This section describes the elements in the Filter window.

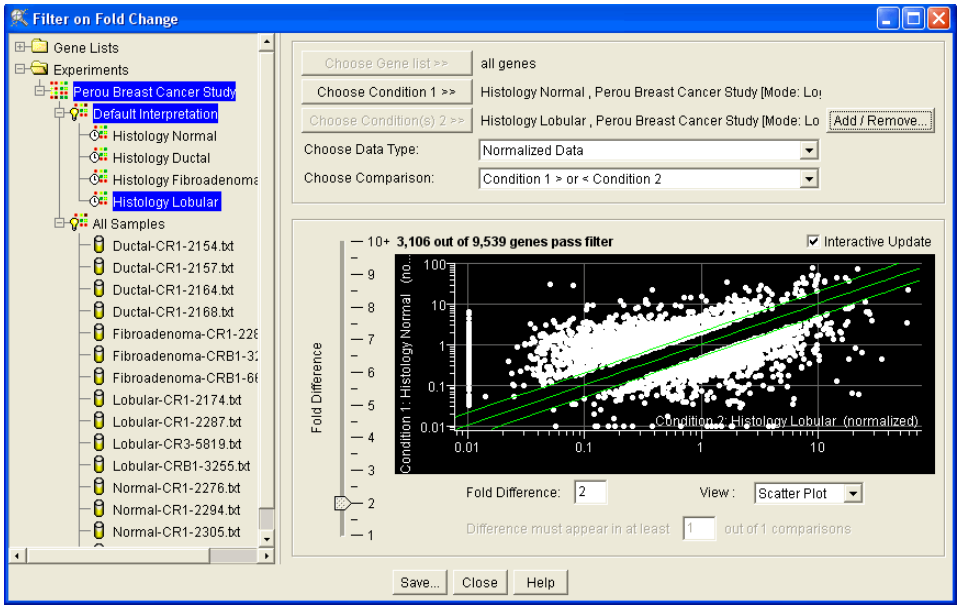


Figure 9-1 Filter Window: Filter on Fold Change

Preview Pane Options

When you open a filtering window, the default view in the Preview pane is based on what type of view makes the most sense for that filtering type. To link the preview display to the main GeneSpring window, select **Main Window** from the **View** menu.

The Preview pane updates dynamically as you change settings for the filter. When working with large experiments, this may cause GeneSpring to respond slowly. To disable this feature, clear the **Interactive Update** box.

Double-ended Sliders

In filters that require you to set a minimum/maximum range, a double-ended slider (Figure 9-2) appears. You can set a range either by using the sliders or by entering numbers directly in the Minimum and Maximum boxes.

You can preserve the size of the specified range while changing the settings by clicking the blue bar between the sliders and dragging it to the desired position.

In some filters, the tics on the slider are not spaced linearly or logarithmically. In these filters, the numbers are spaced so that an equal number of genes fall between each tic. This occurs since using a linear or logarithmic distribution would cause 99% of the genes to fall within three pixels of each other, making the slider impossible to use.

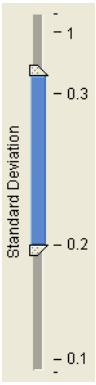


Figure 9-2 Gene List Filter Window: Double-ended Slider

Note: When you enter a number in the Minimum/Maximum box, the slider is moved to that exact number. However, when you move the slider, the number shown in the Minimum/Maximum box is rounded to three digits after the decimal. At the ends of the slider this rounding may sometimes exclude the very biggest or smallest value.

Data Types for Restrictions

You can change the type of data on which to base the restriction, by choosing from a pull-down list in the applicable window. Depending on which feature you are currently using, you may have access to only some of the options in the following list.

- **Normalized Data—**Gene expression values after all normalizations have been applied. These are the default values displayed in various views and are shown in the Normalized column in the Gene Inspector. See "Using the Gene Inspector" on page 7-16 for details.
- **Raw Data—**Experimental data prior to application of any normalizations. This value is used as the numerator to calculate normalized values.

  Note: If your computer's default language is not English, make sure a consistent convention for decimal markers is followed.

- **Control Signal—**A value calculated from all the normalizations applied to the experiment. This value is used as the denominator to calculate normalized values.

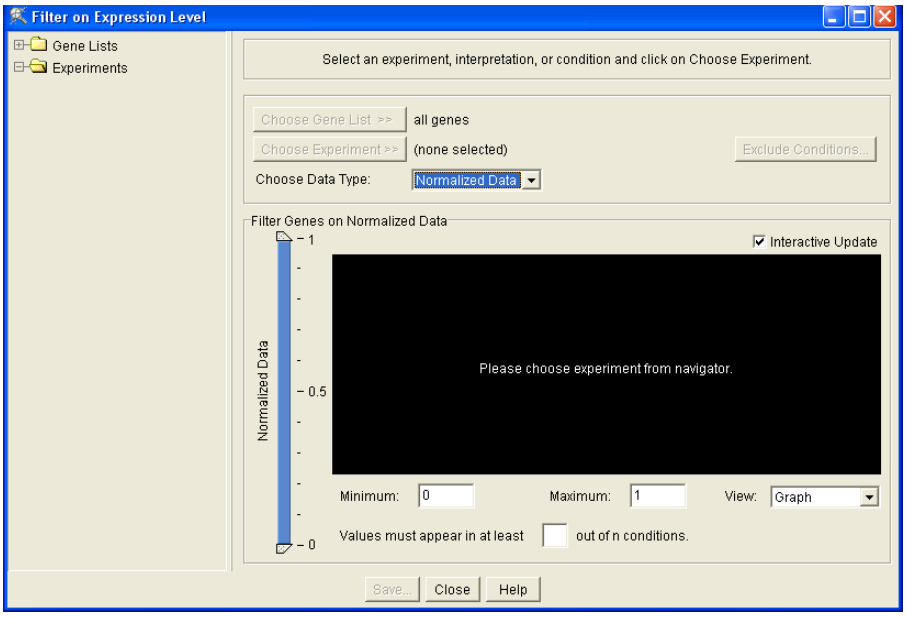
Filtering on Expression Level

Filter on Expression Level finds genes with certain values present in some of the conditions or samples in an experiment or interpretation. You can set what proportion of conditions must meet a certain threshold. For example, to eliminate genes that do not meet a specified control value at least once in the experiment, you can filter them out by setting a minimum expression value to be met in at least one condition.

To filter on expression level:

1. Select **Filtering > Filter on Expression Level**.

   The Filter on Expression Level window opens.



   Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. To choose an experiment to filter, do one of the following:

   - In the **Navigator**, open the **Experiments** folder, right-click the data (experiment, condition, or subset of conditions) you want, and then select the **Set** command.
   - In the **Navigator**, open the **Experiments** folder, select the data (experiment, condition, or subset of conditions) you want, and then click the **Choose Experiment** button.

     By default, the experiment that is currently selected in the GeneSpring window will serve as the starting point for the filter. In most cases this is correct and there is no need to change the experiment. If the selected experiment is not the experiment you want to start the filter with, you can choose another experiment.

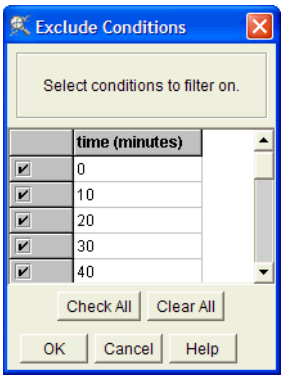
     After you choose an experiment, the name appears in the right panel.

4. In the **Choose Data Type** menu, select the data type you want.

   For more information on data types for filtering, see "Data Types for Restrictions" on page 9-4.

5. To exclude certain samples or conditions from a filter, click the **Exclude Conditions** button.

   The Exclude Conditions window opens. By default, all conditions are selected.



   This window lets you choose the samples or conditions that you do not want to include in the filtering process. This function is useful in situations where it might not be appropriate to include all samples and conditions in the filter. If you do not want to exclude conditions, you can skip this step and use the samples and conditions associated with the experiment you selected in step 3b.

6. Do the following:

   - To exclude a condition, clear the check box.
   - To include a condition, select the check box you want.
   - To include all conditions, click the **Check All** button.
   - To exclude all conditions, click the **Clear All** button.

7. Specify the following values for the filter:

   - **Minimum—**the smallest gene value to allow in your list (also known as the cut-off value).
   - **Maximum—**the largest gene value to allow in your list.
   - **Values must appear in at least [ ] out of [ ] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This value lets you select genes that do not satisfy the filtering selections in all conditions, but only in a subset of the conditions.

     If you want the filter to apply only to certain conditions, use the exclude conditions feature, or use the Advanced Filtering option. Go to "Using Advanced Filters" on page 9-29 for more information.

     Those genes that pass the filter will immediately be shown in the graphical display, unless the "Interactive Update" check box is deselected.

8. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.

9. Go to "Saving Gene Lists" on page 12-6 for more information.

Filtering on Fold Change

Filter on Fold Change finds genes based on a comparison of two samples or conditions. Use this tool to find fold changes in gene expression levels between two samples or conditions.

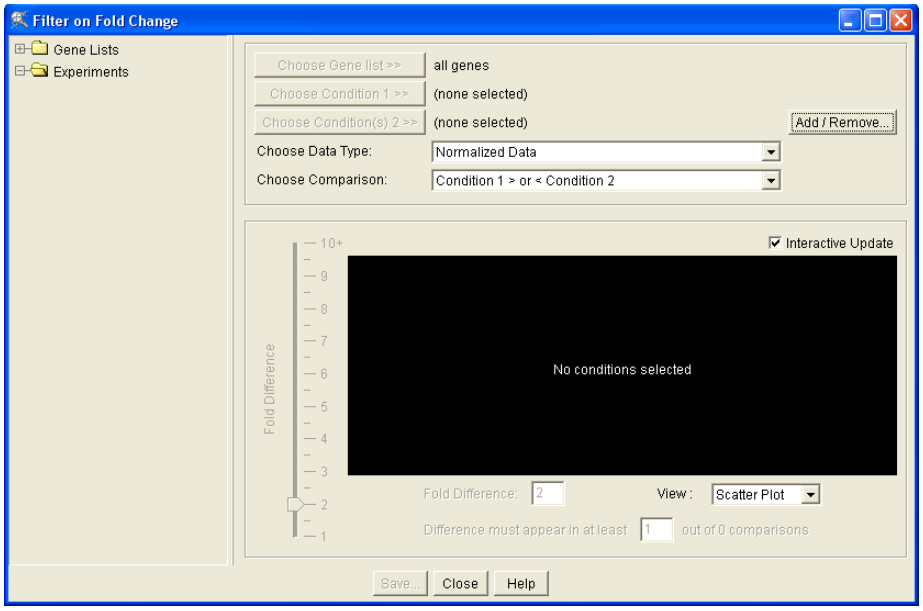
The Fold Change filter will require the definition of at least two different conditions that will be compared. For the fold change calculation, the ratio between Condition 2 and Condition 1 is calculated (Fold change = Condition 2/Condition 1).

Condition 1 should be one, and only one, condition, however, it can also be considered as the base condition. Condition 2 can either be a single condition or a set of conditions. If a single condition is used, the fold change value calculation is straight forward. If multiple conditions are selected for Condition 2, the fold change for each of the conditions in Condition 2 will be calculated.

To filter on fold change:

1. Select **Filtering > Filter on Fold Change**.

   The Filter on Fold Change window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. To choose the conditions, do the following:

   a. In the **Navigator**, open the **Experiments** folder.

   b. To specify a single sample or condition, select it from the **Navigator**, and then click the **Choose Condition(s) 1** button.

   c. To select all the conditions in an experiment, select the experiment in the **Navigator**, and click the **Choose Condition(s) 1** button.

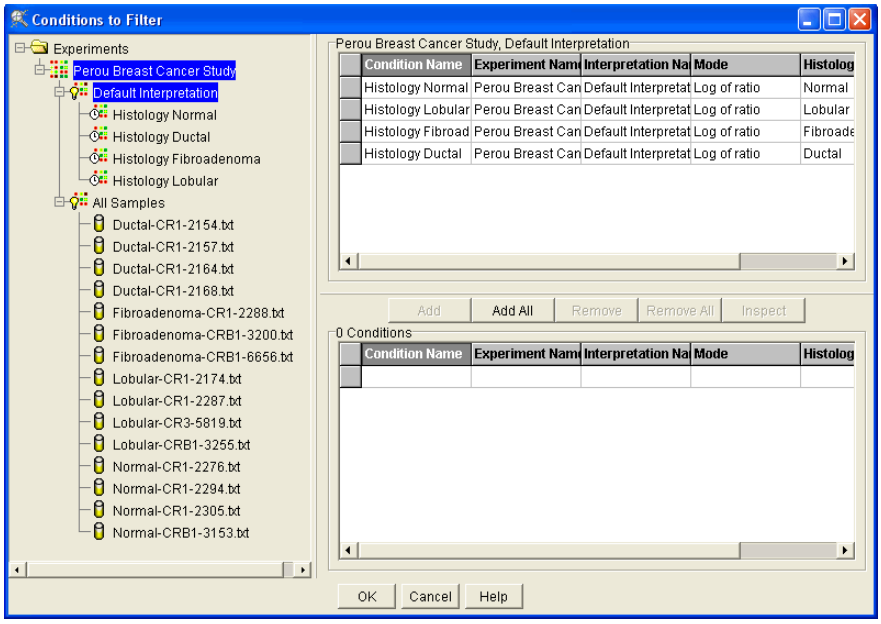
   d. To specify a single sample or condition, select it from the **Navigator**, and then click the **Choose Condition(s) 2** button.

   e. To select all the conditions in an experiment, select the Experiment Interpretation in the **Navigator**, and then click the **Choose Condition(s) 2** button.

   f. To select a pool of conditions manually (from any experiments), click the **Add/ Remove** button.

If more than one condition needs to be selected for Condition 2, but not all of them, you can manually select the conditions to include or exclude using the Add/Remove function.

After choosing the conditions you want, the Conditions to Filter window opens:



4. To specify the conditions to filter, do the following:

   a. In the **Navigator**, open the **Experiment** folder, and then choose an experiment.

      The conditions in that experiment appear in the upper panel to the right of the Navigator.

   b. To add a condition to the filter, select it in the upper panel, and then click the **Add** button.

      The condition is added to the **Selected Conditions** list in the lower panel.

   c. To add all conditions from an experiment, click the **Add All** button.

   d. To remove a selected condition, select it in the lower panel, and then click the **Remove** button.

   e. To remove all selected conditions, click the **Remove All** button.

5. To view a condition in the Condition Inspector, do one of the following:

   - Select the condition in either list and click the **Inspect button.**
   - Double-click the condition.

6. When you are done selecting conditions, click the **OK** button.

7. In the **Choose Data Type** menu, select the data type you want.

   For more information on data types for filtering, see “Data Types for Restrictions” on page 9-4.

8. In the **Choose Comparison** menu, choose whether you want the signal in the first sample or condition to be greater than, less than, equal to, or not equal to (greater than or less than) that in the second sample.
9. Specify a fold factor using the slider, or by entering a value in the **Fold Difference** field.
10. Enter a value in the **Difference must appear in at least [ ] out of [ ] comparisons** field.
11. In the **View** menu, select the option you want for displaying the results. The options are **Main Window** or **Scatter plot**.

    Those genes that pass the filter will immediately be shown in the graphical display, unless the "Interactive Update" check box is deselected.
12. To save genes that pass the filter as a gene list, click the **Save** button.

    The Save Gene List window opens.
13. Go to "Saving Gene Lists" on page 12-6 for more information.

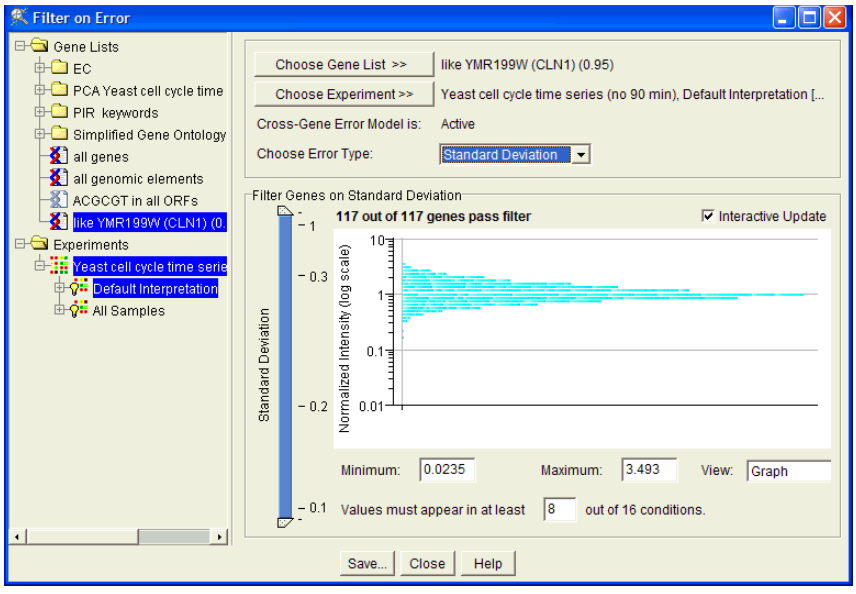
Filtering on Error

Filtering on Error filters on the standard deviation, the standard error, or a range of replicates based on the specified gene list, experiment, or condition.

To filter on errors:

1. Select **Filtering > Filter on Errors**.

   The Filter on Errors window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. To choose an experiment to filter, do one of the following:

   - In the **Navigator**, open the **Experiments** folder, right-click the data (experiment, condition, or subset of conditions) you want, and then select the **Set** command.
   - In the **Navigator**, open the **Experiments** folder, select the data (experiment, condition, or subset of conditions) you want, and then click the **Choose Experiment** button.

     By default, the experiment that is currently selected in the GeneSpring window will serve as the starting point for the filter. In most cases this is correct and there is no need to change the experiment. If the selected experiment is not the experiment you want to start the filter with, you can choose another experiment.

     After you choose an experiment, the name appears in the right panel.

4. In the **Choose Error Type** menu, select the error type to filter on. The options are:

   - **Standard Deviation**—The absolute value of the standard deviation for replicates in the conditions. Use this error type if the number of replicates between conditions is approximately the same.
   - **Standard Error**—The absolute value of the standard error. Use this measure if the number of replicates between conditions varies widely.
   - **Range of Replicates**—The difference between the minimum and maximum value of the replicates in the conditions.

     If you use an experiment that has one condition, the associated value will be the difference between the normalized minimum and maximum. If you use an experiment that has multiple conditions, the associated value will list the number of times the replicate is present.

5. To disable interactive, clear the **Interactive Update** option.

   This option controls the existence of the graph and the X out of Y genes pass filter line. If you clear this option, you will not be given any indication of how many genes pass their filter until you click the **Save** button.

If your experiment is really big, enabling the Interactive Update option may cause your system to slow down. For large experiments, it is recommended that you disable this option.

6. Specify the following values for the filter:
   - **Minimum—**the smallest gene value to allow in your list (also known as the cut-off value).
   - **Maximum—**the largest gene value to allow in your list.
   - **Values must appear in at least [ ] out of [ ] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This value lets you select genes that do not satisfy the filtering selections in all conditions, but only in a subset of the conditions.

     Those genes that pass the filter will immediately be shown in the graphical display, unless the “Interactive Update” check box is deselected.
7. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.
8. Go to “Saving Gene Lists” on page 12-6 for more information.

Filtering on Confidence

Filtering on Confidence filters on either the t-test p-value or the number of replicates based on the specified gene list, experiment, or condition. This filter is useful if you want to identify genes with unreliable measurements and remove them from your experiment. Carrying over poor-quality samples or genes can affect the statistical significance of your clustering analyses findings.

Multiple Testing Correction

When testing a set of genes for statistical significance across various groups, some of the genes may be falsely considered as statistically significant. If 10,000 genes are tested for differential expression between groups, with a significance p-value cutoff of 0.05, then the expected level of genes to be identified as significant by chance alone, even if there is no true differential expression, is 500 genes:

```
10,000 x 0.05 = 500 genes
```

```
Possible false positives = (# of genes) (p-value cutoff)
```

The purpose of a multiple testing correction is to keep the overall error rate/false positives to less than the user-specified p-value cutoff, even if thousands of genes are being analyzed.

If you rely on the nominal p-value when testing the statistical significance of group comparisons for many genes, a significant number of genes pass the filter by chance alone.

For example, if you test 10,000 genes for reliable changes between groups at significance level 0.05, (assuming the tests are independent) you would expect to misidentify about 500 genes as significant, even when there is no real difference in gene expression. Even if you identify 1,000 genes showing significant behavior by this approach, half of the genes on the list appear by chance, which lessens the value of the list. Multiple testing corrections adjust the individual p-value to account for this effect.

To prevent this large number of false positives, you can apply one of a number of multiple testing corrections, such as Bonferroni, Bonferroni Step Down (Holm), or Benjamini and Hochberg False Discovery Rate.

Bonferroni

The Bonferroni multiple testing correction, based on Bonferroni's inequality, limits the chance of a false positive result to be no more than *a* by multiplying each nominal p-value by N (with a maximum of 1). This process controls the FWER, and the expected number of genes by chance is a.

Bonferroni Step Down (Holm)

The Step Down adjustment computes the most significant p-value, and whether it meets the cutoff after multiplying by *N*. If that gene is found to be significant, the next-most significant gene is considered, but the gene that was found significant is removed from the multiple-testing, so the multiple-testing adjustment is now based on *N* - 1. This process is continued as long as genes pass the successive tests. This process controls the FWER, and expected number of genes by chance is *a*.

Benjamini and Hochberg False Discovery Rate

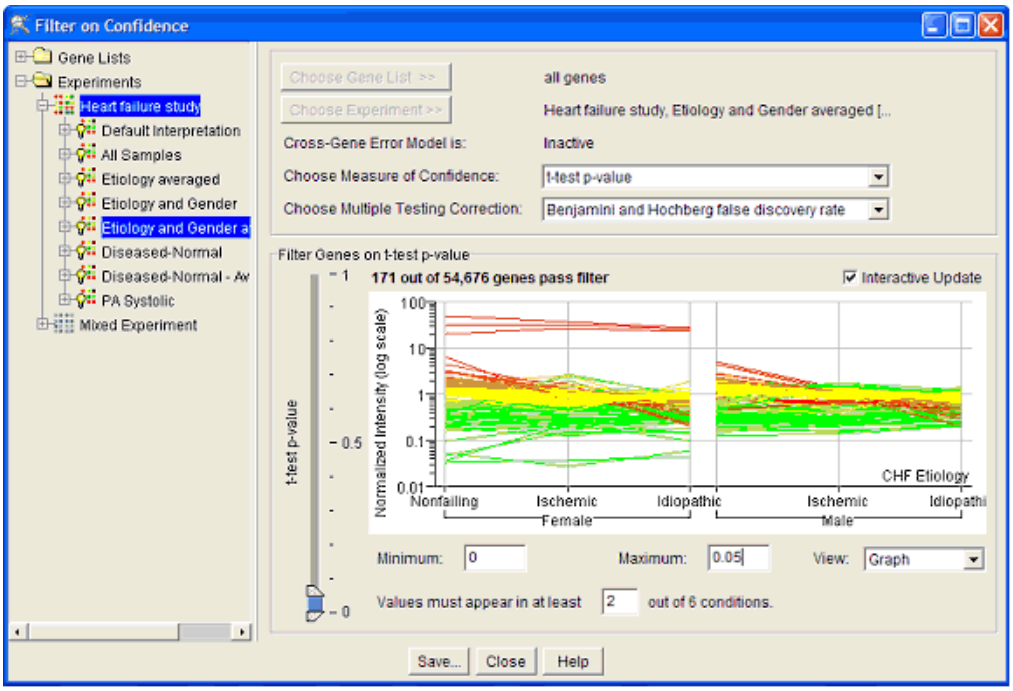
In contrast to the above procedures, the Benjamini and Hochberg procedure controls the false discovery rate (FDR), defined as the proportion of genes expected to occur by chance (assuming genes are independent) relative to the proportion of identified genes. Expected number of genes by chance is a times the number of tests found significant after applying this correction. There is no way to calculate this in advance, so the statement about the number expected simply says expected number of genes by chance is 100α% of the genes identified. This procedure provides a good balance between discovery of significant genes and protection against false positives, since occurrence of the latter is held to a small proportion of the list, and is probably the best choice of multiple-testing correction for most situations.

Filtering on Confidence

To filter on confidence:

1. Select **Filtering > Filter on Confidence**.

   The Filter on Confidence window opens.



   Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. To choose an experiment to filter, do one of the following:

   - In the **Navigator**, open the **Experiments** folder, right-click the data (experiment, condition, or subset of conditions) you want, and then select the **Set** command.
   - In the **Navigator**, open the **Experiments** folder, select the data (experiment, condition, or subset of conditions) you want, and then click the **Choose Experiment** button.

By default, the experiment that is currently selected in the GeneSpring window will serve as the starting point for the filter. In most cases this is correct and there is no need to change the experiment. If the selected experiment is not the experiment you want to start the filter with, you can choose another experiment.

After you choose an experiment, the name appears in the right panel.

4. In the **Measure of Confidence** menu, select a measure of confidence you want. The options are:
   - t-test p-value
   - Number of Replicates

   For details on t-test p-values, see "Using the Gene Inspector" on page 7-16.
5. In the **Choose Multiple Testing Correction** menu, select the correction you want. The options are:
   - Bonferroni
   - Bonferroni Step Down (Holm)
   - Benjamini and Hochberg False Discovery Rate
   - None
6. Specify the following values for the filter:
   - **Minimum—**the smallest gene value to allow in your list (also known as the cut-off value).
   - **Maximum—**the largest gene value to allow in your list.
   - **Values must appear in at least [ ] out of [ ] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This value lets you select genes that do not satisfy the filtering selections in all conditions, but only in a subset of the conditions.
7. In the **View** menu, select the option you want for displaying the results. The options are **Main Window** or **Graph**.

   Those genes that pass the filter will immediately be shown in the graphical display, unless the "Interactive Update" check box is deselected.
8. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.
9. Go to "Saving Gene Lists" on page 12-6 for more information.

Filtering on Parameters

Filter on Parameter calculates the correlation between expression values and parameter or attribute values. This filter lets you find genes that show some correlation with any of the experiment parameters or sample attributes. For example, if one of the sample attributes is LDL level, you could find those genes that show some correlation in their expression value to the LDL value. This filter only works for numerical parameters or attributes.

Parameters

Experiment parameters are variables that describe the individual samples in a manner that is important for the analysis of the expression data. Experiment parameters are defined for each of the samples of an experiment.

Attributes

GeneSpring handles attributes included in the Filter on Parameter option as follows:

- If an attribute does not have a value for every sample in the selected experiment, only those samples that have values for the attribute are used in the calculation.
- If an attribute has non-numeric values for some samples, those samples are treated as if they have no values.
- If an attribute has different units, such as seconds vs. minutes, then each of the different attribute and unit pairs are treated as completely separate attributes.
- The selected attribute or parameter values define the conditions for the purpose of the calculation.

Note: The attributes that you choose to include in this analysis should all be on the same scale of measurement. For example, all the Age attribute values should start counting from the same point, rather then comparing Age from conception numbers to Age from birth numbers.

Replicates

Replicates can be multiple spots on the same array representing the same gene (also referred to as a copy), the same sample in more than one array or a biological replicate - that is equivalent samples taken from more than one organism.

A parameter defined as a replicate is graphically a hidden variable; no visual distinction is made based upon this parameter or its parameter values.

Replicates are averaged on a log scale. When comparing expression values to parameter or attribute values, GeneSpring uses a ratio scale.

Similarity Measures

To use the Filter on Parameter tool, you need to select a similarity measure to use. The equation used to determine the overall correlation is as follows:

$$X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}$$

Table 9-1 describes the variables used in this equation.

Table 9-1 Correlation Equation Variables

A	The correlation coefficient between the gene in question in experiment 1 and the selected gene, also from experiment 1.
a	The weight specified for experiment 1.
B	The correlation coefficient of the gene in question in experiment 2, to the selected gene, also from experiment 2.
b	b is the weight associated with experiment 2.
C	The correlation coefficient of the gene in question in experiment 3 to the selected gene, also from experiment 3.
c	The weight associated with experiment 3.

Table 9-2 describes the different similarity measure options that you can use in the Filter on Parameter tool.

Table 9-2 Similarity Measures

Similarity Measure	Description
Standard Correlation	Measures the angular separation of expression vectors for Genes **A** and **B** around zero. Result = $\mathbf{a}.\mathbf{b}/(\lvert\mathbf{a}\rvert\lvert\mathbf{b}\rvert)$
Smooth Correlation	Makes a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph window. Do the same to make a vector **B** from **b**. *Result* = $\mathbf{A}.\mathbf{B}/(\lvert\mathbf{A}\rvert\lvert\mathbf{B}\rvert)$
Change Correlation	Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i)-\pi/4$.Do the same to make a vector **B** from **b**. *Result* = $\mathbf{A}.\mathbf{B}/(\lvert\mathbf{A}\rvert\lvert\mathbf{B}\rvert)$

Table 9-2 Similarity Measures (Continued)

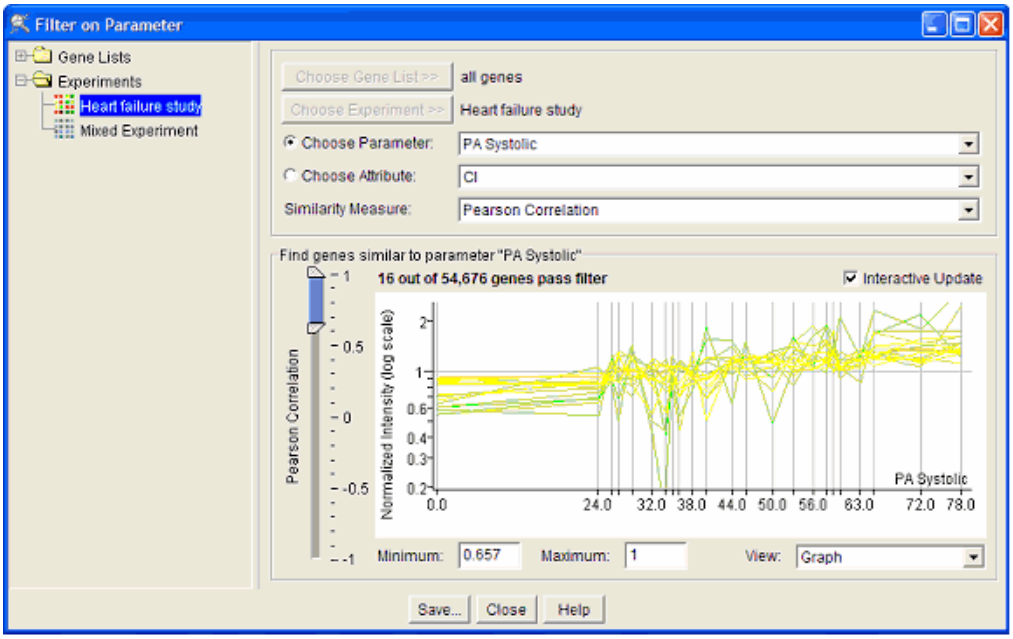
Upregulated Correlation	Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is max(atan(a_{i+1}/a_i)-π/4,0). Do the same to make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)
Pearson Correlation	Calculate the mean of all elements in vector **a**. Then subtract that value from each element in **a**. Call the resulting vector **A**. Do the same for **b** to make a vector **B**. *Result* = **A.B**/(|**A**||**B**|)
Spearman Correlation	Order all the elements of vector **a**. Use this order to assign a rank to each element of **a**. Make a new vector **a**' where the i^{th} element in **a**' is the rank of a_i in **a**. Now make a vector **A** from **a**' in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)
Spearman Confidence	Compute a value *r* of the spearman correlation as described above. *Result* =1-(probability you would get a value of *r* or higher by chance.)
Two-sided Spearman Confidence	Compute a value *r* of the Spearman correlation as described above. *Result* =1-(probability you would get a value of |*r*| or higher, or -|r| or lower, by chance.)

Filtering on Parameters

To filter on parameter:

1. Select **Filtering > Filter on Parameter**.

   The Filter on Parameter window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. To choose an experiment to filter, do one of the following:

   - In the **Navigator**, open the **Experiments** folder, right-click the data (experiment, condition, or subset of conditions) you want, and then select the **Set** command.
   - In the **Navigator**, open the **Experiments** folder, select the data (experiment, condition, or subset of conditions) you want, and then click the **Choose Experiment** button.

     By default, the experiment that is currently selected in the GeneSpring window will serve as the starting point for the filter. In most cases this is correct and there is no need to change the experiment. If the selected experiment is not the experiment you want to start the filter with, you can choose another experiment.

     After you choose an experiment, the name appears in the right panel.

4. Select the **Choose Parameter** or **Choose Attribute** option and select the option you want from the list.

   You can filter either on Experiment Parameters or Sample Attributes. The parameter or attribute you filter on must be a numeric. Go to "Experiment Parameters" on page 5-37 and "Sample Attributes" on page 5-58 for more information.

5. In the **Similarity Measure** menu, select the measure you want. The options are:

   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation
   - Spearman Confidence
   - Two-sided Spearman Confidence

6. Select the standard correlation you want by using the **Standard Correlation** slider.

   The numbers on the slider correspond to the correlation chosen in th**e Similarity Measure** menu. The values range from 0 (no correlation), 1 (a perfect positive correlation), or -1 (a perfect negative correlation).

7. In the **View** menu, select the view you want. The options are:

   - **Cumulative Distribution**—shows a graph of the number of genes that pass the cutoff against the correlation coefficient.
   - **Graph**—sets the display as the graph view, allowing you to see the expression values of the genes that pass the filter.
   - **Main Window**—sets the view to whatever the view is in the Genome Browser.

8. To hide the view, clear the **Interactive Update** option.

   This option controls the graph and the X out of Y genes that pass the filter line. If you clear this option, you will not be given any indication of how many genes pass their filter until you click the **Save** button. This option is recommended for large sets of data since the update may take some time.

9. Specify the following values for the filter:

   - **Minimum—**the smallest gene value to allow in your list (also known as the cut-off value).
   - **Maximum—**the largest gene value to allow in your list.
   - **Values must appear in at least [ ] out of [ ] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This value lets you select genes that do not satisfy the filtering selections in all conditions, but only in a subset of the conditions.

Those genes that pass the filter will immediately be shown in the graphical display, unless the “Interactive Update” check box is deselected.

10. To save genes that pass the filter as a gene list, click the **Save** button.

    The Save Gene List window opens.

11. Go to “Saving Gene Lists” on page 12-6 for more information.

Filtering on Flags

GeneSpring lets you find genes based on the data quality flags in the original data files. Flags are additional measurement markers in your data set. There are four possible flag values “Present” or “P”, “Absent” or “A”, “Marginal” or “M” and “Unknown”.

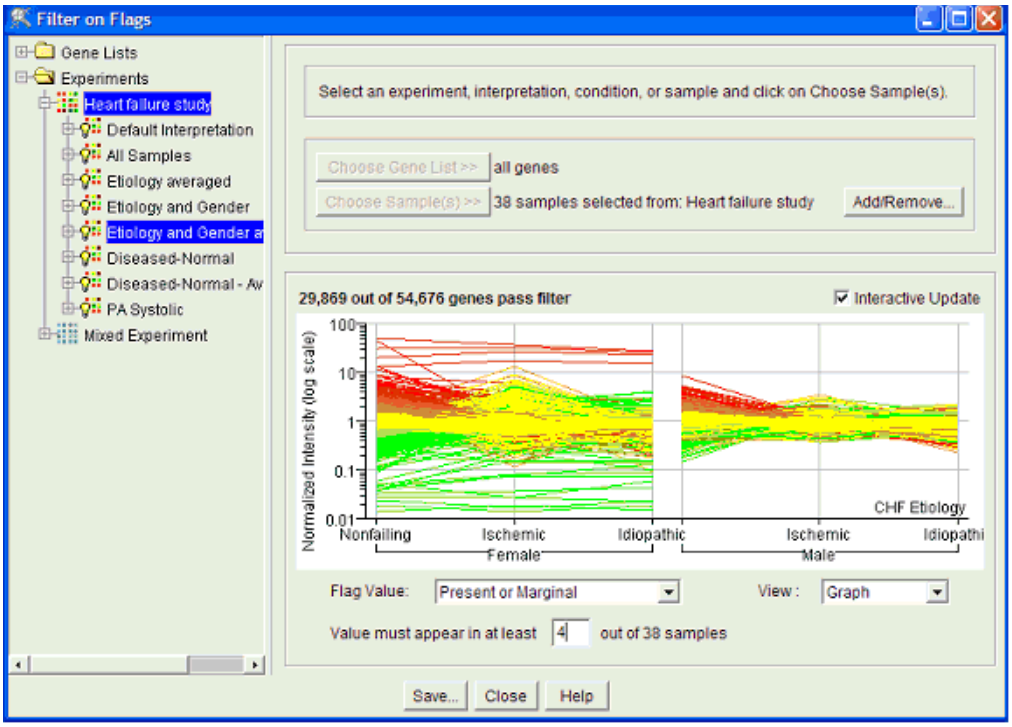
These Flag values are assigned during import of the data sets and the A, P and M calls are special codes in GeneSpring. Although these codes are the same ones used in the Affymetrix platform, it is possible to map other types of flags to this set of four flags. For more information about setting up flags, go to “Setting Up Columns” on page 5-10.

The Filtering on Flags option is available only if a flag column was specified in the data file when it was loaded into GeneSpring.

To filter on flags:

1. Select **Filtering > Filter on Flags**.

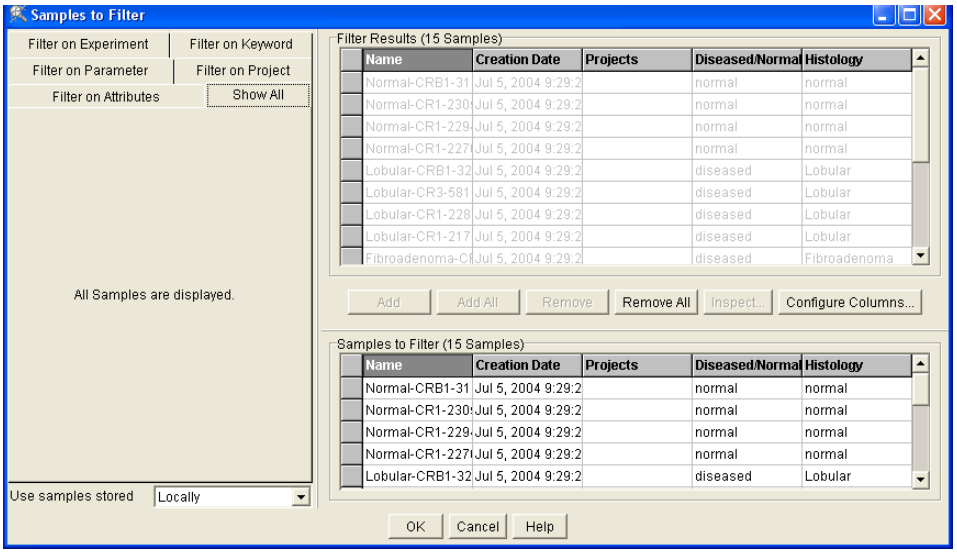
   The Filter on Flags window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. Do one of the following:
   - To use all the samples from an experiment, select an experiment from the **Navigator** and click the **Choose Samples** button.
   - To select individual samples from the selected experiment, click the **Add/Remove** button.

   The Samples to Filter window opens.



This window provides the same functions as the Sample Manager window. For more information about the Sample Manager, go to "Sample Manager Window" on page 5-45.

3. In the **Flag Value** menu, select the flag value you want. The options are:
   - Anything
   - Present
   - Present or Marginal
   - Present or Unknown
   - Present, Marginal, or Unknown
   - Marginal
   - Absent
   - Unknown
4. Enter a value in the **Difference must appear in at least [ ] out of [ ] samples** field.

5. In the **View** menu, select the option you want for displaying the results. The options are **Main Window** or **Graph**.

   Those genes that pass the filter will immediately be shown in the graphical display, unless the “Interactive Update” check box is deselected.

6. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.

7. Go to “Saving Gene Lists” on page 12-6 for more information.

Filtering on Data Files

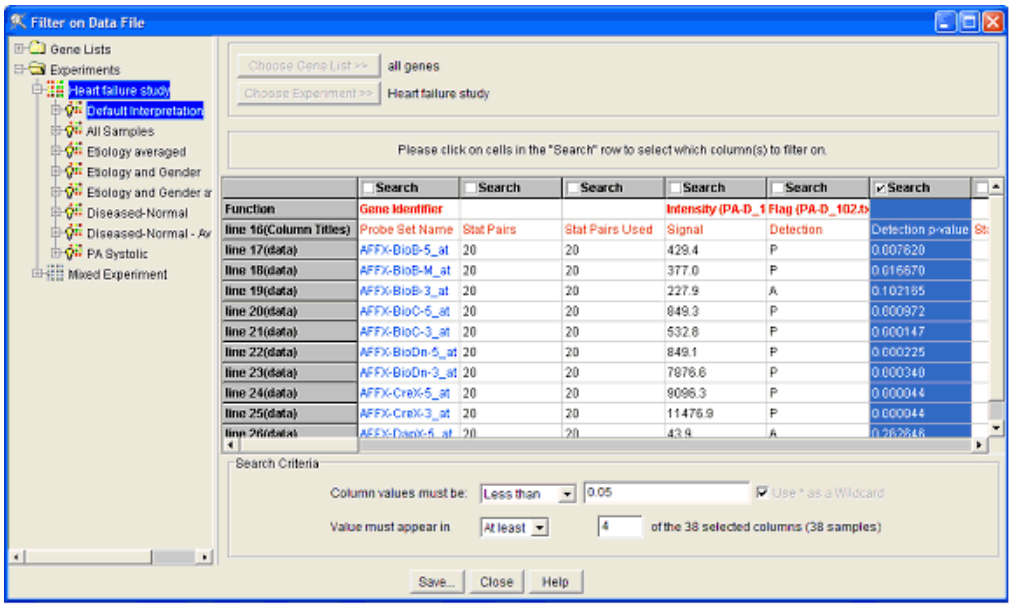
Filter on Data File lets you filter genes based on values in a specific column of your original data files. For example, if your data file contains some data types that are not loaded directly into GeneSpring (like the detection p-value in the Affymetrix CHP files or % Saturation in the GPR file), you can still filter this information with the Filter on Data file filter. This filter lets you choose any of the columns in your data file and filter on the contents, both numeric and character data.

If your sample data files are in multiple formats, this window opens with a separate tab for each data format. The available options on each tab are the same as the options for the standard Data File Restrictions window.

To filter on a data file:

1. Select **Filtering > Filter on Data File**.

   The Filter on Data File window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:
   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.
3. To choose an experiment to filter, do one of the following:
   - In the **Navigator**, open the **Experiments** folder, right-click the data (experiment, condition, or subset of conditions) you want, and then select the **Set** command.
   - In the **Navigator**, open the **Experiments** folder, select the data (experiment, condition, or subset of conditions) you want, and then click the **Choose Experiment** button.

     By default, the experiment that is currently selected in the GeneSpring window will serve as the starting point for the filter. In most cases this is correct and there is no need to change the experiment. If the selected experiment is not the experiment you want to start the filter with, you can choose another experiment.

     After you choose an experiment, the name appears in the right panel.
4. To select the column or columns to search on, select the **Search** box in the header of the columns in your experiment.

   The column is highlighted in yellow.
5. In the **Search Criteria** box, do the following:
   a. To restrict column values, select a value from the **Column Values Must Be** menu. The options are:
      - Less than
      - Greater than
      - Equal to
      - Not equal to
      - Contain

   b. Enter the search term in the field provided.
   c. To perform a wildcard search, add an asterisk character in the search term, and select the **Use * as Wildcard** option.

      For example: to find zero or more characters, excluding spaces and punctuation enter an asterisk (*). *g*n* finds words such as gene and genome.

d. In the **Value must appear** menu, select the number of columns in which the values must appear and enter a value in the field provided.

   The number to the right of this box indicates the number of columns that have been selected. If you have multiple data formats, this number reflects the total number of columns selected on all tabs.

   Those genes that pass the filter will immediately be shown in the graphical display, unless the “Interactive Update” check box is deselected.

6. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.

7. Go to “Saving Gene Lists” on page 12-6 for more information.

Filtering on Arbitrary Files

Filtering on Arbitrary File lets you find genes based on the information in one or more columns from a selected file. You can perform only one search at a time. The selected file must have at least two columns: a column for gene identifiers and a column of some other type of data. The **Match Gene Identifier To** menu lets you specify which type of term you are using to identify each gene.

For instance, if you chose **Systematic Name**, **Common Name**, or **Synonym**, GeneSpring looks for the specified identifier in any of those three columns in any of your master table of genes files. The filter then returns a list of the genes that have matching identifiers in the selected field and pass the filter in the **Search Criteria** fields.

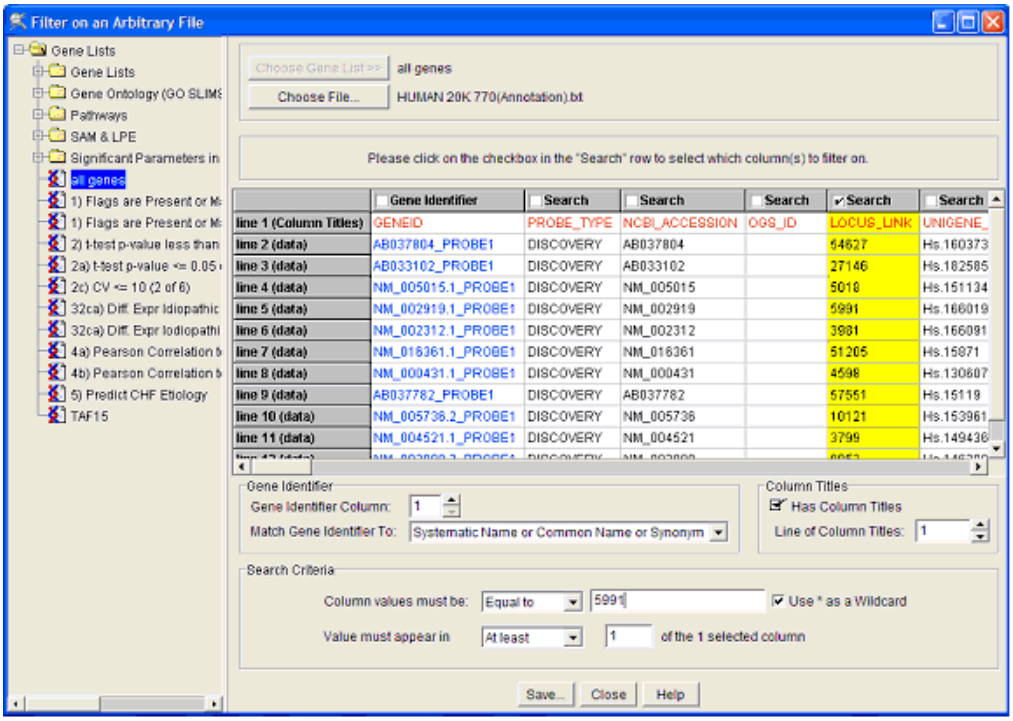
The same search criteria are applied to every selected column; therefore all selected columns must contain the same type of information.

To perform multiple searches of columns containing different information, you must apply multiple restrictions to your data, one for each type of information.

To filter on an arbitrary file:

1. Select **Filtering > Filter on Arbitrary File**.

   The Filter on an Arbitrary File window opens.



Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. Click the **Choose File** button.

   The Select One File window opens.

3. Select the file you want from the browse menu. This file must have a column of gene identifiers.

   GeneSpring loads the file and the file appears in a table in the window. During the loading process, GeneSpring analyzes the file to determine which column contains the Gene Identifier, and colors that column in blue. In addition, it attempts to guess whether the file has column titles, and colors that row red.

4. If GeneSpring did not select the correct column for the gene identifier, specify it in the **Column Containing Gene Identifier** field.

5. Use the **Match Gene Identifier To** menu to specify the column to which the gene identifier should be matched.

6. If the column header row chosen is incorrect, use the **First Line of Data** field to adjust the number of rows. If GeneSpring did not identify any column header row, you must first check the **Has Column Titles** box.

7. To select the column or columns to search on, select the **Search** box in the header of the columns in your experiment.

The column is highlighted in yellow.

8. In the **Search Criteria** box, do the following:

   a. To restrict column values, select a value from the **Column Values Must Be** menu. The options are:

      - Less than
      - Greater than
      - Equal to
      - Not equal to
      - Contain

      For example, if you load an Affymetrix file, use the menu to select the **Abs/call** column and search for all entries equal to "M". This produces a list of only marginal data.

   b. Enter the search term in the field provided.

   c. To perform a wildcard search, add an asterisk character in the search term, and select the **Use * as Wildcard** option.

      For example: to find zero or more characters, excluding spaces and punctuation enter an asterisk (*). *g*n* finds words such as gene and genome.

   d. In the **Value must appear** menu, select the number of columns in which the values must appear and enter a value in the field provided.

      The non-editable number to the right of this box indicates the number of columns that have been selected. If you have multiple data formats, this number reflects the total number of columns selected on all tabs.

   Those genes that pass the filter will immediately be shown in the graphical display, unless the "Interactive Update" check box is deselected.

9. To save genes that pass the filter as a gene list, click the **Save** button.

   The Save Gene List window opens.

10. Go to "Saving Gene Lists" on page 12-6 for more information.

Filtering on Gene List Numbers

GeneSpring can filter genes according to the numbers associated with them in a gene list. When you make a new list based on a filter or similarity metric, the value used as a filter is associated with the genes on the new list.

Some examples of associated numbers are correlation coefficients, p-values, fold change ratios, or in the case of a regulatory sequence search, the number of base pairs before the promoter region. Associated numbers can be found by double-clicking a gene list to bring up the Gene List Inspector.

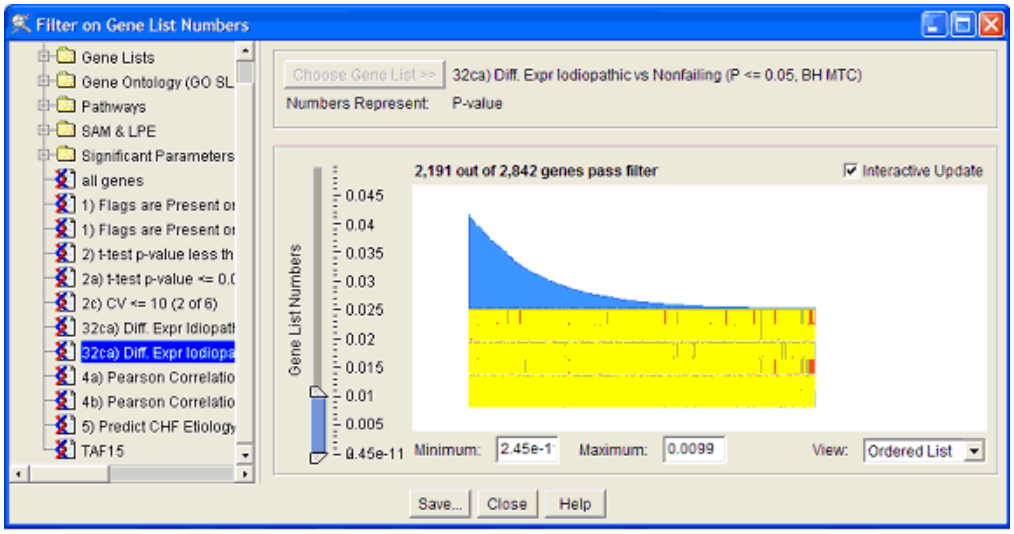
Filtering genes by their associated numbers is helpful if you want to use this information to create a more specific list of genes. For example, you may want to find genes that are very similar to another gene (with a high correlation coefficient), or genes that are a

specific distance from a promoter found using the Find Potential Regulatory Sequences tool. For details, go to "Finding Potential Regulatory Sequences" on page 14-1.

To filter on gene list numbers:

1. Select **Filtering > Filter on Gene List Numbers**.

   The Filter on Gene List Numbers window opens.



   Filtering is typically a multi-step process in which the result of the first filter becomes input for subsequent filters. By default, GeneSpring will select the currently selected gene list as the starting gene list. Only those genes that are members of the selected gene list will be subjected to the filter. In most cases, this is the desired behavior. If the gene list that is selected by default is not the correct gene list, you can choose a different gene list to start the filtering process.

2. To choose a gene list, do one of the following:

   - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Set** command.
   - In the **Navigator**, open the **Gene Lists** folder, select the gene list you want, and then click the **Choose Gene List** button.

     The name of the gene list appears in the right panel.

3. Use the double-ended slider or enter minimum and maximum restriction values in the fields provided.

   If a gene list has no associated numbers, you cannot select it for filtering in this view. For example, this filter cannot be applied to the "all genes" or "all genomic elements" lists because there are no associated values.

References

Benjamini, Y. and Hochberg, Y. (1995) “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society B*, 57, 289 -300.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2000) “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”. *Department of Statistics Technical Report* #578, University of California, Berkeley (http://stat-ftp.berkeley.edu/tech-reports/index.html)

Holm, S. (1979) “A Simple Sequentially Rejective Bonferroni Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65 -70.

Miller, R.G. (1981) *Simultaneous Statistical Inference*, Second Edition. New York: Springer-Verlag.

Westfall, P.H. and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods*

Using Advanced Filters

The Advanced Filtering window (Figure 9-3) lets you combine basic filters and analysis filters to perform more complex filtering operations. All of the basic filters described in the previous section are available, as well as Filter on Gene List and Filter on Annotations filters. Advanced filtering tools enable you to pass data through multiple filters. They also let you use files that are external to GeneSpring, but that contain important information about the genes.

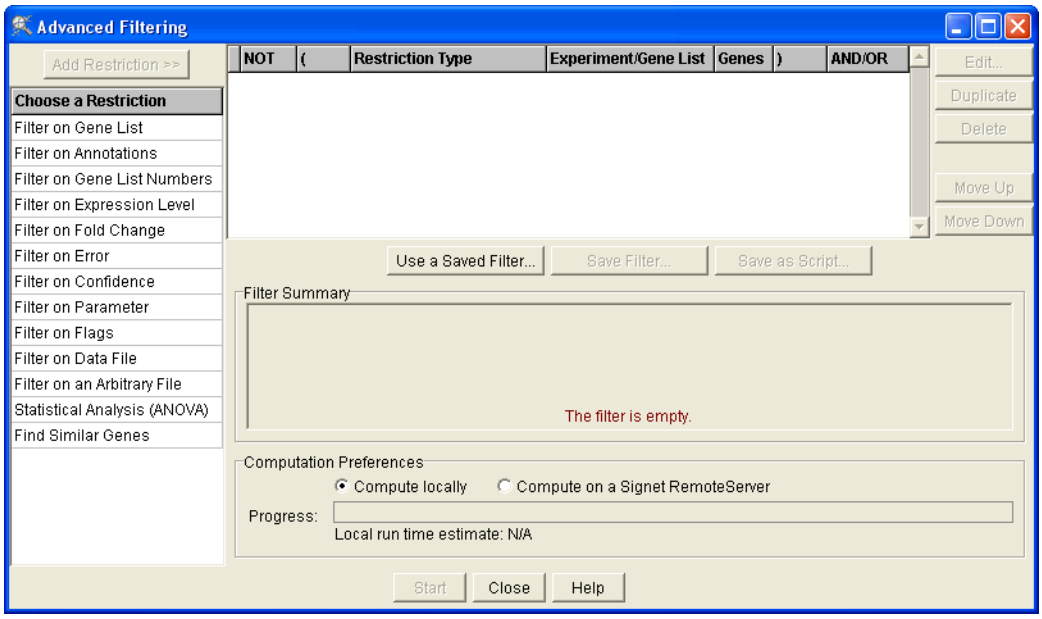


Figure 9-3 Advanced Filtering Window

After performing a filtering operation, you can perform Statistical Analysis (ANOVA) and Find Similar Genes operations with the resulting data. For more information on these operations, go to "Statistical Analysis (ANOVA)" on page 13-1 and "Inspecting Gene Lists" on page 12-11.

Restrictions List

The Restrictions list contains the different filters that you can choose. The options are:

- Filter on Gene List
- Filter on Annotations
- Filter on Gene List Numbers
- Filter on Expression Profile
- Filter on Fold Change
- Filter on Error
- Filter on Confidence
- Filter on Parameter
- Filter on Flags
- Filter on Data File
- Filter on Arbitrary File
- Statistical Analysis (ANOVA)
- Find Similar Genes

Restrictions Table

The end result of the filtering process is a gene list. The Restrictions table shows each of the filtering steps in the process as one row. The default Boolean operator for each of the steps is the AND statement, which means that the genes that are saved at the end of the filtering, will satisfy ALL of the criteria. You can change the Boolean operators to OR or NOT and use parentheses to make more complex filtering statement.

There are no priorities between statements, so without parentheses to group statements, the order is assumed to be left-to-right. To set priorities in filtering, you can create a Boolean filter.

Buttons

- **Add Restriction**—Adds the selected filter to lists of restrictions.
- **Use a Saved Filter**—Lets you select a previously saved filter to use.
- **Save Filter**—Saves the current filter for future use. Saving the filter with this option, allows the script to be recalled at a later date, run again, or edited.
- **Save as Script**—Saves the filter as a script. This option creates a script in the main Scripts folder that can be run and edited like any other script. Saving a complex filter as a script is also a good way to learn how to use the scripting tool.

- **Start**—Starts the filtering process.
- **Close**—Closes the window.
- **Help**—Displays online Help.

Creating Advanced Filters

To create an advanced filter:

1. Select **Filtering > Advanced Filters**.

   The Advanced Filter window opens.

2. In the **Choose a Restriction** list, do one of the following:

   - Select a filter from the list of available filters and click the **Add Restriction** button.
   - Double-click the name of the filter you want.

   An Advanced Filtering window opens.

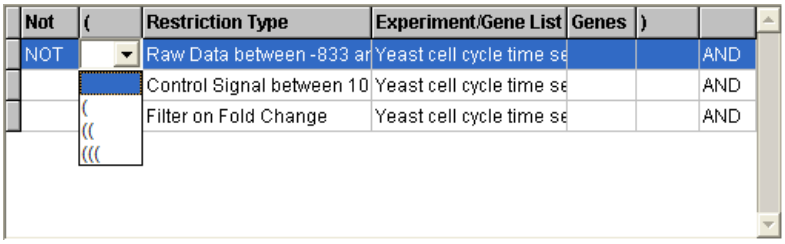
   Note: Each entry in the Advanced Filtering window must result in one and only one gene list. As a result, you do not have the option to run Post Hoc testing when applying 1-way ANOVA in an advanced filter, and you must select an individual gene list of interest when applying 2-way ANOVA.

3. Select the filtering parameters you want and click the **OK** button.

   A line appears representing your filter.

4. Repeat steps 2 and 3 to add additional filters.
5. To re-order steps, select a filtering step and click **Move Up** or **Move Down**.
6. To insert another instance of a filtering step, select the step, and click the **Duplicate** button.
7. To remove a filtering step, select the step and click the **Delete** button.
8. To edit an individual filtering step, do the following:

   - In the **Restrictions** table, double-click the filtering step, and edit it.



   - Select a filtering step, click the **Edit** button, and then edit the step.

9. To create a Boolean filter, use the menus in the **Restrictions** table to group the restrictions together.

   - Select **AND/OR** to tell GeneSpring how to combine the grouped restrictions.

- Select **NOT** to tell GeneSpring not to use the genes from the selected filtering step. Use the parentheses to group filter steps together. You can use a maximum of 3 levels of nesting with the parentheses, which should be enough for most situations.

10. In the **Computation Preferences** box, specify whether to run the script locally or on a remote execution server.

    This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

    If you specify Remote Execution, the Preview pane in the Advanced Filtering window is disabled. This action prevents GeneSpring from executing the filter in real-time while you are creating it.

    Note: An advanced filter using the Arbitrary File Restriction filter cannot be executed remotely.

11. Click the **Start** button.

    As the filter executes, you can watch its progress in the Progress bar.

12. To save the filter, click the **Save Filter** button or the **Save As Script** button.

    For more information on saving filters, see "Saving Advanced Filters" on page 9-32.

Saving Advanced Filters

You can save your filter either as a saved filter or as a script.

Save Filter Option

Saved filters include all of the inputs to the filter and any associated information, including each restriction and its settings. They can be accessed in any genome.

When a saved filter is opened in a genome other than the one in which it was created, the genome data objects (such as gene lists or experiments) appear blank or undefined. You must define these fields within the new genome before running the filter. Filters that require data objects to be defined are displayed in red in the Restrictions table.

A saved filter is limited to the computer on which it was created. It cannot be sent to another user.

Save As Script Option

Filters saved as scripts save all of the current inputs as default inputs, but those inputs are not required to run the script. The script is saved in the Scripts folder in GeneSpring's Navigator. It will not reconstruct the appearance of the Advanced Filtering window; instead, it runs exactly like a standard GeneSpring script. For more information on scripts in GeneSpring, see "Working with Scripts" on page 16-1.

When you save a filter as a script, it is not limited to the computer on which it was created. This means you can send it to other GeneSpring users.

Filtering Data Objects Assigned to Projects

Each data object in GeneSpring can be assigned to one or more projects. A project can be envisaged as a virtual container of data objects that are all part of the same project. A project can contain data objects from one or more genomes and is not limited to the data objects from one genome, allowing the data objects from different arrays to be combined into one project.

GeneSpring enables you to easily search for data objects that have been assigned to projects. It also enables you to see all the data objects that are assigned to a project, regardless of the genome with which they are associated.

Filtering Assigned and Unassigned Projects with the Show Menu

To limit the number of data objects that are shown in the Navigator, you can display only those objects that are assigned to one or more projects. This feature will enable you to more easily find those data objects that are relevant to the current project.

You can use the Show menu (Figure 9-4) in the GeneSpring window to filter assigned and unassigned projects.

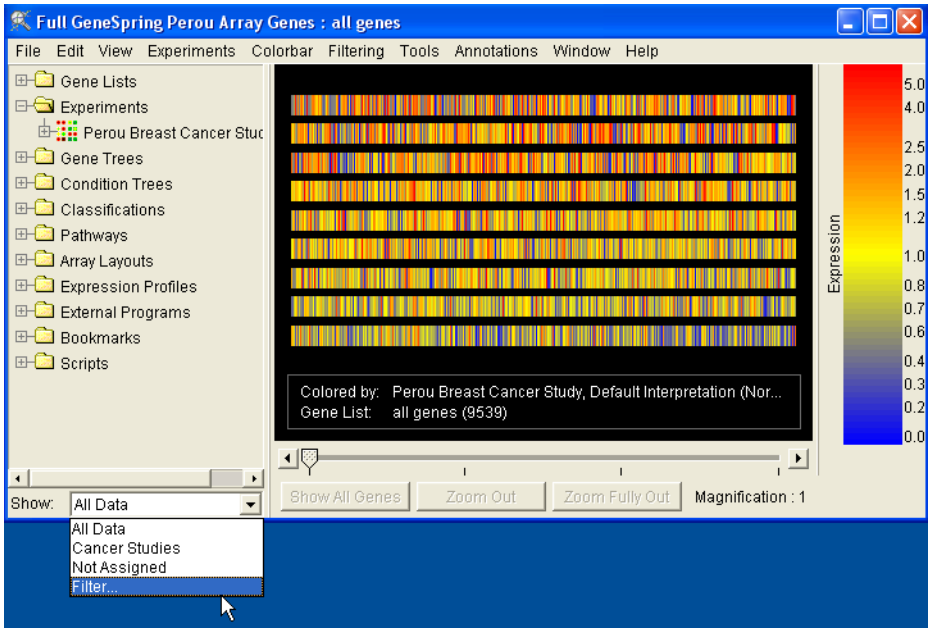


Figure 9-4 GeneSpring Main Window: Show Menu

To filter assigned or unassigned projects:

- In the **Show** menu, do one of the following:
  - Select the project name.
  - Select the **Not Assigned** command.

Filtering Projects Assigned to Data Objects

The Filter Navigator window (Figure 9-5) lets you create custom search filter to locate specific data objects that appear in the Navigator. Go to "Filter Navigator Window" on page 7-41 for more information about this window.

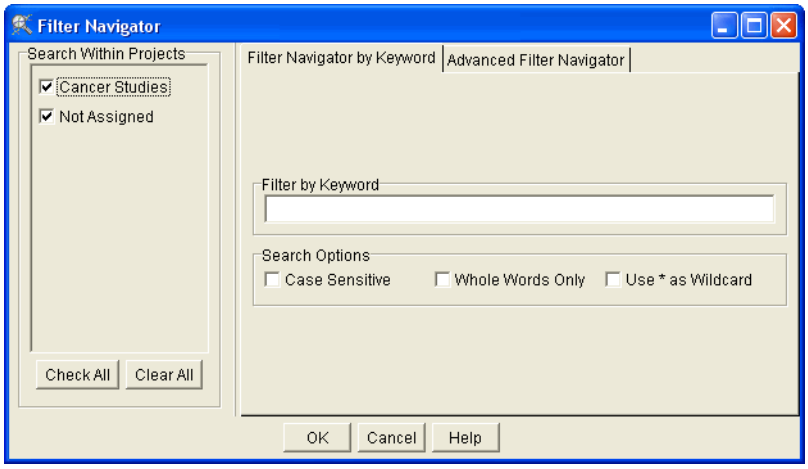


Figure 9-5 Filter Navigator Window

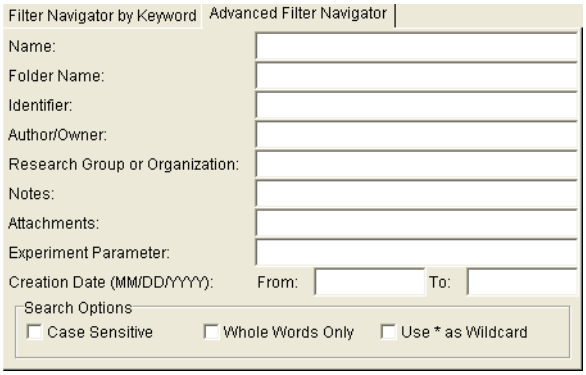
To filter projects assigned to data objects

1. To open the **Filter Navigator**, do one of the following:
   - Select **Edit > Filter Navigator**.
   - Select the **Show** menu and then choose the **Filter** command.

   The Filter Navigator window opens.
2. To filter projects by keyword, do the following:
   a. Enter the search criteria that you want.

      Go to "Filter Navigator Window" on page 7-41 for complete information.

   b. Click the **OK** button.
3. To perform advanced filtering, do the following:
   a. Select the **Advanced Filter Navigator** tab.



b. Enter the search criteria you want.

   Go to "Filter Navigator Window" on page 7-41 for complete information.

c. Click the **OK** button.

Results are displayed in the Search Results window.

10

Normalizing Data

This chapter describes the different strategies that are available to normalize your experiment data. It covers the following topics:

- Experiment Normalizations
- Data Transformations
- Normalization Methods
- Normalization Strategies for Specific Technologies

Experiment Normalizations

Experiment normalizations are used to standardize your microarray data to enable differentiation between real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales your data so that you can compare relative gene expression levels.

Sixteen transformations are available for creating powerful and flexible normalization scenarios. Normalization steps can be applied in virtually any order, and include operations such as dye swapping experiments and median polishing. Scenarios can be saved and applied in other experiments.

GeneSpring provides the following data transformation and normalization options.

- SAGE Transformation
- Real Time PCR Transformation
- Subtract Background Based on Negative Controls
- Set Measurements less than 0.01 to 0.01
- Dye Swap
- Per-spot Normalization
- Per-chip Normalizations
- Per-gene Normalizations

This chapter describes how to use these different options.

Normalizing Experiments

This section explains how to normalize experiments. It covers the following topics.

- Experiment Normalizations Window
- Assumptions
- Normalization Methods
- Opening the Experiment Normalizations Window
- Adding Normalization Steps
- Normalizing Specific Samples
- Editing Normalization Steps
- Removing Normalization Steps
- Reordering Normalization Steps
- Applying Default Normalizations
- Viewing Detailed Descriptions of Normalizations
- Saving a Normalization Scenario
- Working with Saved Scenarios
- Normalization Warnings

Experiment Normalizations Window

The Experiment Normalizations window (Figure 10-1) lists the normalizations currently applied to your experiment and lets you add, edit, delete, or re-order normalization steps. You can save the current normalization steps as a scenario for future use, or load a previously saved scenario. The Warnings panel displays information about requirements or other potential problems with the currently specified normalizations.

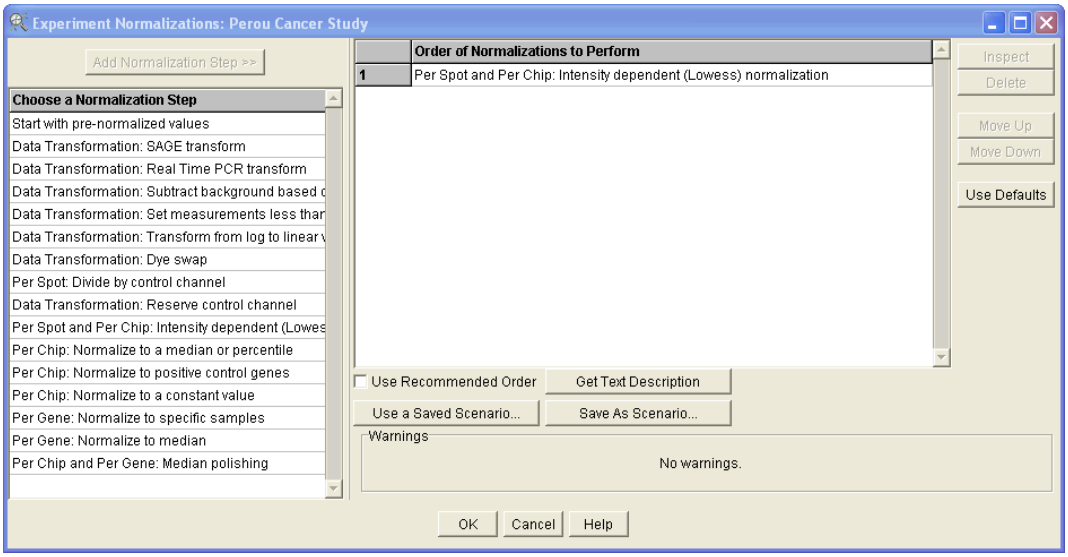

Experiment Normalizations: Perou Cancer Study
Add Normalization Step >>
Choose a Normalization Step
Start with pre-normalized values
Data Transformation: SAGE transform
Data Transformation: Real Time PCR transform
Data Transformation: Subtract background based
Data Transformation: Set measurements less than
Data Transformation: Transform from log to linear
Data Transformation: Dye swap
Per Spot: Divide by control channel
Data Transformation: Reserve control channel
Per Spot and Per Chip: Intensity dependent (Lowes
Per Chip: Normalize to a median or percentile
Per Chip: Normalize to positive control genes
Per Chip: Normalize to a constant value
Per Gene: Normalize to specific samples
Per Gene: Normalize to median
Per Chip and Per Gene: Median polishing
Order of Normalizations to Perform
1
Per Spot and Per Chip: Intensity dependent (Lowess) normalization
Inspect
Delete
Move Up
Move Down
Use Defaults
Use Recommended Order
Get Text Description
Use a Saved Scenario...
Save As Scenario...
Warnings
No warnings.
OK
Cancel
Help

Figure 10-1 Experiment Normalizations Window

Assumptions

GeneSpring assumes the data you have entered is raw data and must be normalized. If your data has been pre-normalized around a median other than 1, it may not be accurately interpreted during analysis. If your data is pre-normalized this way, see "Normalize to a Constant Value" on page 10-19.

Normalization Methods

There are several ways to normalize your data in GeneSpring. Typically, you will want to do either one per-chip normalization together with one per-gene normalization, or one per-spot normalization together with one per-chip normalization. There are important exceptions to these recommendations, however, which are discussed in this section.

Opening the Experiment Normalizations Window

To open the Experiment Normalizations window:

- Select **Experiments > Experiment Normalizations**.

Adding Normalization Steps

To add a normalization step:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Choose a Normalization Step** table, select the normalization step you want to add by either clicking the **Normalization Step** button or double-clicking the selected step.
4. Do one of the following:
   - To add a per-spot normalization, go to "Per-spot Normalization" on page 10-12.
   - To add a per-chip normalization step, go to "Per-chip Normalizations" on page 10-15.
   - To add a per-gene normalization step, go to "Per-gene Normalizations" on page 10-20.
   - To use a normalization strategy for a specific technology, go to "Normalization Strategies for Specific Technologies" on page 10-24.
5. Click the **OK** button to add this step to your normalization scenario.

Normalizing Specific Samples

Most normalizations can be applied in any order, and different samples in the same experiment can be normalized in different ways. You have the option of applying most normalization steps only to specific samples in your experiment.

To normalize specific samples:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Choose a Normalization Step** table, select the normalization step you want to add by either clicking the **Normalization Step** button or double-clicking the selected step.
4. Select the **Apply Only to Specific Samples** box.

   A list of samples in the current experiment appears. If these samples are named, the names appear as sample identifiers. By default, each sample is named for the file it is from. If there is more than one sample in a file, the column name can also be included.

   Note: Samples that cannot be normalized; for example, samples with no normalized column, appear grayed out in the list and cannot be selected.

5. Select the box of any samples to which you want to apply this step.
6. Click the **OK** button to add this step to your normalization scenario.

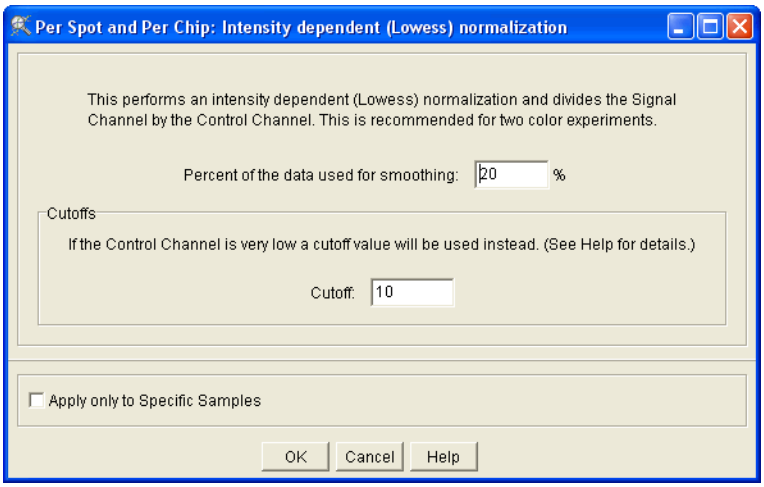
Editing Normalization Steps

To edit a normalization step:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Order of Normalizations to Perform** table, select the step you want to edit.
4. Do one of the following:
   - Double-click the step number.
   - Select the step and click the **Inspect** button.

     The Configuration window for the selected normalization step appears.


Per Spot and Per Chip: Intensity dependent (Lowess) normalization
This performs an intensity dependent (Lowess) normalization and divides the Signal Channel by the Control Channel. This is recommended for two color experiments.
Percent of the data used for smoothing:
20
%
Cutoffs
If the Control Channel is very low a cutoff value will be used instead. (See Help for details.)
Cutoff:
10
Apply only to Specific Samples
OK
Cancel
Help

5. Edit the normalization settings, as necessary.
6. Click the **OK** button.

Removing Normalization Steps

To remove a normalization step:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Order of Normalizations to Perform** table, select the step you want to delete.

   Be certain you have selected the correct step, since no confirmation dialog opens.
4. Click the **Delete** button.
5. Click the **OK** button.

Reordering Normalization Steps

To reorder a normalization step:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Order of Normalizations to Perform** table, select the step you want to reorder.
4. Click **Move Up** or **Move Down**, as necessary, to reorder the step.
5. Click the **OK** button.

Applying Default Normalizations

When an experiment is created during the sample import process, normalizations are applied before you reach the Experiment Normalization window. These normalizations are determined from the data format of the samples in the experiment. For information on the default normalizations used during sample import, see “Using the Default Normalizations” on page 5-35.

When you create an experiment from samples that have already been imported, the default normalizations are the generic one-color and generic two-color scenarios. These normalizations can be applied when you create an experiment using the Create New Experiment menu or the Experiment Normalizations window.

The following procedure applies default normalizations to your data. It also removes any normalization changes you have applied to your data and applies only the default settings.

To apply default normalizations:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. Click the **Use Defaults** button.
4. Click the **OK** button.

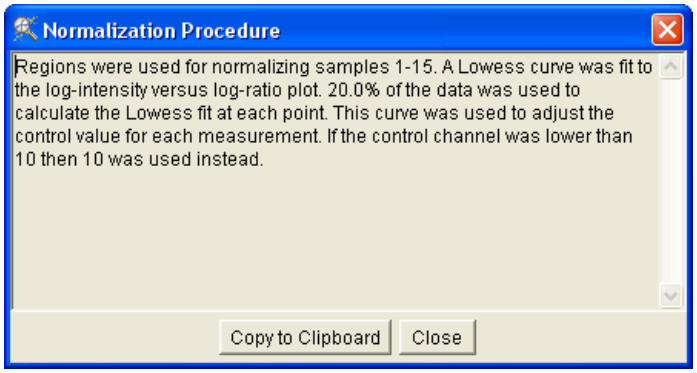
Viewing Detailed Descriptions of Normalizations

To view a more detailed description of a normalization:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Choose a Normalization Step** table or the **Order of Normalizations to Perform** table, select the step you want.
4. Click the **Get Text Description** button.

   A window opens with a description of the selected normalization step.


Normalization Procedure
Regions were used for normalizing samples 1-15. A Lowess curve was fit to the log-intensity versus log-ratio plot. 20.0% of the data was used to calculate the Lowess fit at each point. This curve was used to adjust the control value for each measurement. If the control channel was lower than 10 then 10 was used instead.
Copy to Clipboard
Close

5. To copy the text in this window to the keyboard, click the **Copy to Clipboard** button.

   You can then paste the text into a text editor.

Saving a Normalization Scenario

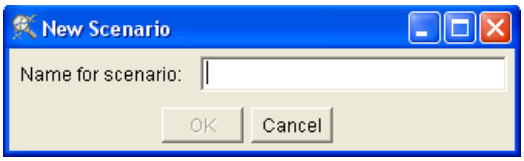
You can save the current normalization sequence for use in other experiments. A saved scenario records whether each step was applied to all samples or just a limited number of samples. It does not, however, record to which samples the steps were applied. Additionally, it does not record a list of positive or negative controls or a list of control samples (as in the Normalize to Specific Samples option).

To save a normalization scenario:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. In the **Choose a Normalization Step** table, select the normalization step you want to add by either clicking the **Normalization Step** button or double-clicking the selected step.
4. Click the **Save As Scenario** button.

   The New Scenario window opens.



5. In the **Name for scenario** field, enter the name you want.
6. Click the **OK** button.

   Once you save a normalization scenario, it is available for use in all experiments.

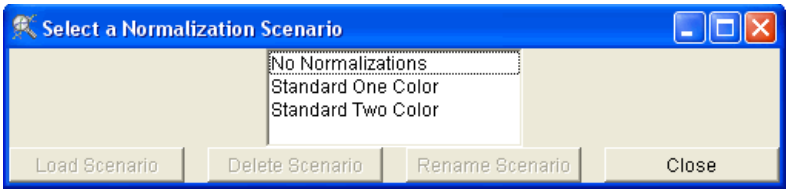
Working with Saved Scenarios

To work with a saved scenario:

1. In the **Navigator**, open the **Experiments** folder, and then select the experiment you want.
2. Select **Experiments > Experiment Normalization**.

   The Experiment Normalization window opens.
3. Click the **Use a Saved Scenario** button.

The Select a Normalization Scenario window opens.



4. Do one of the following:
   - To select a scenario from the list and use it in the current experiment, select the scenario you want, and then click the **Load Scenario** button.
   - To remove a scenario from the list, select the scenario you want, and then click the **Delete Scenario** button.
   - To rename a scenario in the list, select the scenario you want, and then click the **Rename Scenario** button.
5. Click the **Close** button.

Normalization Warnings

Normalization warnings can occur under the following circumstances:

- A normalization step is missing
- A normalization step is inappropriate (for example, there are too few genes or samples)
- Normalizations are applied to only some of the samples

Warnings appear in orange. You can proceed with an active warning, but the results may not be what was intended. Fatal errors appear in red. A fatal error means that the current normalization steps will not produce a usable result. In this case, the OK button is disabled until the problem is solved.

When a warning or error applies to a specific normalization step, that step is displayed in the list in the appropriate color for the warning or error.

Data Transformations

Data transformation is the process of applying a mathematical modifications to the values of a variable. When you choose a data transformation method, GeneSpring recalculates the data values and uses them in any subsequent analyses that you perform.

This section describes the transformations that can be applied to your data. These data transformation types include the following:

- SAGE Transformation
- Real Time PCR Transformation
- Subtract Background Based on Negative Controls
- Set Measurements less than 0.01 to 0.01
- Dye Swap

SAGE Transformation

The SAGE transformation method is recommended only for SAGE data. It fills in zeroes for all genes not mentioned in your data file.

Real Time PCR Transformation

The Real Time PCR transformation method, doubling measurements are converted into measurements of mRNA concentration using the equation 2^{-n}.

Subtract Background Based on Negative Controls

In the Subtract Background transformation method, the median value of the gene list is subtracted from the raw values for each gene. The gene list used can be typed in or loaded from a file.

Typing in Gene Lists

To type in a gene list:

1. Enter a gene in each line of the text box provided.
2. Right-click in this box to use the **Copy** and **Paste** options.

Loading Gene Lists

To load a gene list from a file:

- Click **Load From File** and select the gene list from the browse window that appears.

If there are already genes listed when you click **Load From File**, the genes from the list you select are added to the existing list of genes. Any genes you have already entered are not overwritten.

Note: This text box can contain no more than 32,000 text characters (including carriage returns).

The list of negative control genes should be intersected with the regions, if there are any. Negative controls should be averaged within each region. If any region does not have any negative controls, an error message appears alerting you that the normalization cannot be performed.

Set Measurements less than 0.01 to 0.01

The Set Measurements option sets any measurements less than a specified cutoff value to the cutoff value. By default, this value is 0.01. To enter another value, click in the Cutoff text box and enter a new value. This step can be applied before or after other normalizations.

Transform from Log to Linear Values

The Log to Linear Values option transforms logarithmic data into linear expression values. This is required if your raw data are reported in log values, since GeneSpring requires data to be linear. To view your data on a logarithmic scale, use the experiment interpretation. For more information on experiment interpretations, see “Setting Up Experiment Interpretations” on page 5-65.

To apply this normalization, specify the base of the original measurements by selecting the appropriate option. The available options are:

- Base 2
- Natural Log (e)
- Base 10
- Other— Enter a base value in the provided text box

Dye Swap

The Dye Swap option swaps your control channel and signal channel to compare dyes in two-color experiments.

This step must be the first normalization step that changes the units of signal and control. For example, a log transform can precede this step, but it cannot precede a per-chip normalization. This is because the signal and control must be in the same units.

Normalization Methods

This section describes the different methods of normalizations that can be applied to data. These types include the following:

- Start with Pre-Normalized Values
- Per-spot Normalization
- Per-chip Normalizations
- Per-gene Normalizations

Start with Pre-Normalized Values

The Start with Pre-normalized Values option is provided for backwards compatibility and lets you maintain normalizations from a previous experiment. It can be applied only to samples that were created by GeneSpring's Merge/Split window or uploaded before GeNet 3.0. If you select this option when none of the data in your experiment have been previously normalized, a message alerting you to do this is displayed and the **OK** button is disabled.

Per-spot Normalization

This section describes the options that are available for per-spot normalizations. These options are:

- Divide by Control Channel
- Reserve Control Channel
- Intensity-dependent Normalization

Per-spot normalizations are commonly used for two-color experiments. The formula for this normalization is:

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(control channel value for gene A in sample X)}}$$

Divide by Control Channel

The Divide by Control Channel option divides the measured intensity of each gene by the value of its control channel. This is recommended for two-color experiments if you do not use intensity-dependent normalization. If the control channel value is very low, a cutoff value is used instead. By default, this value is 10.0. To change this value, click in the Cutoff text box and enter the desired cutoff value.

Note: The cutoff value cannot be lower than 0.000001.

This normalization works as follows:

	Signal >= Cutoff	Signal < Cutoff
Control >= Cutoff	Signal/Control	Signal/Control
Control < Cutoff	Signal/Cutoff	No Data

Reserve Control Channel

The Reserve Control Channel option was previously known as Use Control Channel for Trust. This option tells GeneSpring to use the control channel to determine the saturation of the color of your genes. This option is recommended when you import the signal to control ratio and the control channel.

Enter the value below which you do not trust the control signal in the Cutoff text box. By default, this value is 10.0.

Intensity-dependent Normalization

Intensity-dependent normalization (often called non-linear or LOWESS normalization) is recommended for use in most two-color experiments. This step can be applied only to chips with more than 100 genes. LOWESS normalization uses region designators in the same way that other per-chip normalization methods do. For details, see “Region Normalization” on page 10-19.

Intensity dependent normalization is a technique that is used to eliminate dye-related artifacts in two-color experiments that cause the Cy5/Cy3 ratio to be affected by the total intensity of the spot.

This normalization process attempts to correct for artifacts caused by non-linear rates of dye incorporation as well as inconsistencies in the relative fluorescence intensity between some red and green dyes. Such artifacts often result in a curve in the graph of raw versus control signal (see panel A in Figure 10-2).

In the absence of bias, one would expect there to be no dependence of raw signal on control signal and thus the data points would be scattered symmetrically around the 45^o line.

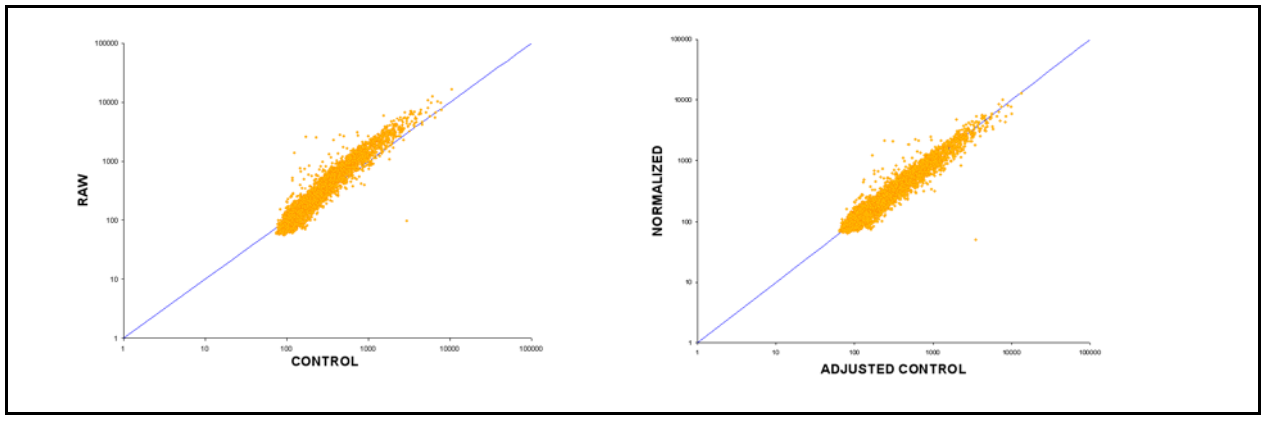


Figure 10-2 Effect of Intensity-dependent Normalization

GeneSpring's intensity-dependent normalization feature fits a curve through the data and uses this curve to adjust the control value for each measurement. When the resulting normalized data are graphed versus the adjusted control value, the points are distributed more symmetrically around the 45^o line (see Figure 10-2, panel B). You can specify the percent of data to be used for smoothing. By default, this value is 20.0%.

To counter the problem of taking logarithms of negative values (in subsequent steps), and to discount outliers, the raw and control values for each spot (R, G) are shifted by a constant.

$$shift = max\left(\frac{median(R)}{100} - min(R), \frac{median(G)}{100} - min(G), 0\right)$$

The data goes through the following transformations:

$$\begin{aligned}
&(G,R) \longrightarrow (G - shift, R - shift)\\
&\longrightarrow (\log(G - shift), \log(R - shift))\\
&\xrightarrow{L} (A, M)\\
&\longrightarrow (A, M - f(A))\\
&\xrightarrow{L^{-1}} (\log(G'), \log(R'))\\
&\longrightarrow (G', R')\\
&\longrightarrow (G' + shift, R' + shift)\\
&\longrightarrow \left(\frac{R(G' + shift)}{R' + shift}, R\right)
\end{aligned}$$

where $f(A)$ is the fitted function of the transformed data, and the coordinates are transformed by the operation:

$$L = \begin{bmatrix} 1/2 & 1/2 \\ -1 & 1 \end{bmatrix}$$

The last step of the transformation is performed because normalizations in GeneSpring are accomplished by adjusting the control value and leaving the raw value unchanged.

The fit of the data, $f(A)$ is made using the LOWESS algorithm where the value f = 0.2 is used for the fraction of the total data points used for smoothing at each point (see "References" on page 10-29 for more information on the LOWESS algorithm).

The degree of the polynomial fitted is 1. For efficiency, the regression is not calculated at each data point, but at a progressively fitted mesh that adjusts to the sparsity of the data.

If you attempt to re-normalize an experiment that has been constructed using the Merge/Split Experiment tools, you will be unable to apply intensity dependent normalization. The control values in merged experiments have already been adjusted and thus do not reflect the intensity of the reference dye.

If you perform an intensity-dependent normalization, it is usually not necessary to perform a per chip normalization. Like normalizing to the distribution of all genes, *intensity-dependent normalization should not be used on specialized arrays that contain a small number of genes, or on arrays where a majority of the genes may react similarly to experimental conditions*.

Per-chip Normalizations

Per-chip normalizations control for chip-wide variations in intensity. Such variations may be due to inconsistent washing, inconsistent sample preparation, or other microarray production or microfluidics imperfections. GeneSpring does not allow you to perform more than one per chip normalization, as they all address the same issue.

There is no dedicated option for region normalization. However, if you have region designators, all per-chip normalizations are performed on each region independently.

This section describes the options that are available for per-chip normalizations. These options are:

- Normalize to a Median or Percentile
- Normalize to Positive Control Genes
- Normalize to a Constant Value
- Region Normalization
- Affine Background Correction
- Divide by Specific Samples

Normalize to a Median or Percentile

The Normalize to a Median or Percentile option lets you divide all of the measurements on each chip by a specified percentile value. By default, this value is 50.0%. To change this value, enter a new one in the text box. You do not have to restrict the measurements used in the calculation of the percentile. You can limit measurements based on a specified cutoff or by flag values.

Limiting Measurements by Flag Values

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction.

To limit measurements by flag values:

1. Select the **Use only measurements flagged** box.
2. Select the appropriate option from the menu. The available options are:
   - Present Only
   - Present or Marginal
   - Anything but Absent

Limiting Measurements by a Cutoff Value

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units. The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file.

The cutoff value is re-calculated, based on the previously applied normalization steps, in most cases the normalization to the median of the chip. The calculation is as follows:

The new gene median cutoff = the currently set raw cutoff value (10) / the median of the set of median chip values

To limit measurements by a cutoff value:

1. Select the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the menu.
3. Enter the cutoff figure in the text box.

   The default value is 10.0.

Applying Background Correction

To apply background correction:

- Select the appropriate box in the Background Correction section of the window. You have the following options:
  - **Never apply extra background correction.**
  - **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes.

- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Global Per-chip normalization is not recommended in any experiment where more than 50% of the genes on the chip are likely to be affected similarly by the experimental conditions. For example, if a chip containing only known growth factors were used to study differential expression in malignant and benign tumors, you might expect a majority of the genes to be differentially expressed. In this case, applying a per-chip normalization would mask the changes in expression.

Normalize to Positive Control Genes

Some chips come with positive controls (mRNA from another genome or *housekeeping genes*), which are used to control for differences in the amount of exposure between samples. The formula for this difference is:

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(median signal of the positive controls in sample X)}}$$

To normalize to positive control genes, first enter a list of genes. This gene list can be typed in or loaded from a file.

Typing in Gene Lists

To type in a gene list:

1. Enter a gene in each line of the text box provided.
2. Right-click in this box to use the **Copy** and **Paste** options.

Loading Gene Lists

To load a gene list from a file:

1. Click **Load From File** and select the gene list from the browse window that appears.

   If there are already genes listed when you click **Load From File**, the genes from the list you select are added to the existing list of genes. Any genes you have already entered are not overwritten.

   Note: This text box can contain no more than 32,000 text characters (including carriage returns).

2. Select the percentile of the positive controls by which to divide each sample. By default, this value is 50.0%.

Limiting Measurements by Flag Values

You can limit measurements based on a specified cutoff or by flag values. If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction.

To limit measurements by flag values:

1. Select the **Use only measurements flagged** box.
2. Select the appropriate option from the menu. The available options are:
   - Present Only
   - Present or Marginal
   - Anything but Absent

Limiting Measurements by a Cutoff Value

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

The Partially Normalized option means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit measurements by a cutoff value:

1. Select the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

Applying Background Correction

To apply background correction:

- Select the appropriate box in the Background Correction section of the window. The options are:
  - **Never apply extra background correction**.
  - **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes.
  - **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Normalize to a Constant Value

If you are using a technology that calculates its own number for normalization, you will want to use constant values. For instance, Affymetrix's Global Scaling™ centers your data around 2500; in this case you would need to normalize your data to 2500 to center it around 1.

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(hard number in sample X)}}$$

To normalize to a constant value, simply enter the desired value in the **Per Chip: Normalize to a constant value** text box. By default, this value is set to 1.0.

Region Normalization

Regions are assigned in the column editor during the data loading process. For more information, see "Using the Column Editor" on page 5-5. If you have defined regions in your data, all normalization steps are applied on a per-region basis.

There are three ways of designating regions:

- Each data file for a sample is assumed to be a separate region
- Each distinct value in the Region Column is designated as a region
- A specified list of region codes (which may or may not be suffixes in the region column). This option is included for backwards compatibility.

Affine Background Correction

If negative values form a large fraction of your data set, GeneSpring may perform an affine background correction. If a large percentage of your data are negative, normalization can be a problem. For instance, the median, which GeneSpring divides your data by in **Use Distribution of All Genes**, can be very small or even negative.

In such cases, GeneSpring readjusts the background level for your data by adding a constant to all raw control strengths such that the 10th percentile is set equal to 0. The affine background correction is applied only when the 10th percentile is more negative than the median of the data are positive. If the correction is applied, a warning message appears during data loading. Also, in the Gene Inspector, control strengths adjusted by this correction are flagged with asterisks.

To specify your choice for background correction:

- Select the appropriate box in the Background Correction section of the window. You have the following options:
  - **Never apply extra background correction**.
  - **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes.

- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Per-gene Normalizations

This section describes the options that are available for per-gene normalizations. These options are:

- Divide by Specific Samples
- Normalize to Median
- Median Polishing

Divide by Specific Samples

In the Divide by Specific Samples option, each gene is divided by the intensity of that gene in a specific control sample or by the average intensity in several control samples. The formula for this is:

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(signal strength of gene A in the control sample[s])}}$$

Or,

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(average signal strength of gene A in several control sample[s])}}$$

Figure 10-3 shows an example of per-gene normalization to specific samples.

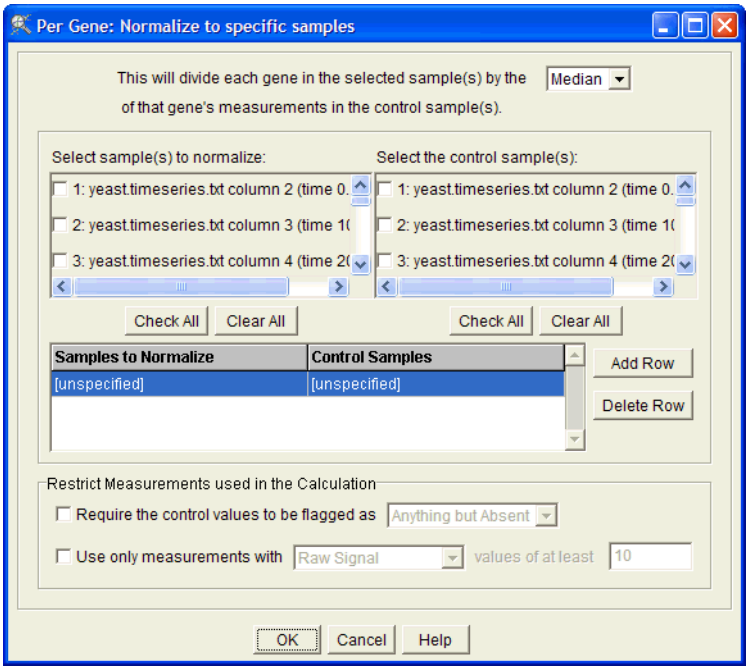


Figure 10-3 Per-gene Normalization: Normalize to Specific Samples

Dividing by Specific Samples

To divide by specific samples:

1. Specify whether to divide by the mean or median by selecting from the menu at the top of the window.
2. Choose a pair of samples and their control samples by selecting the boxes next to the desired samples or typing them into the table.

   Copy and Paste functions will also work in this table. Use commas or semicolons to separate samples, or dashes to indicate a range of samples.
3. To select all samples in a column, click the **Check All** button.
4. To clear all samples in a column, click the **Clear All**.
5. To define a new pair, use the **Add Row** button.
6. To remove a pair, use the **Delete Row** button.

Limiting Measurements by Flag Values

You can limit measurements based on a specified cutoff or by flag values. If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction.

To limit measurements by flag values:

1. Select the **Use only measurements flagged** box.
2. Select the appropriate option from the menu. The available options are:
   - Present Only
   - Present or Marginal
   - Anything but Absent

Limiting Measurements by a Cutoff Value

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

The Partially Normalized option means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit measurements by a cutoff value:

1. Select the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

Note: You cannot perform this normalization and normalize to the median of each gene, because they address the same issue.

Normalize to Median

The Normalize to Median option accounts for the difference in detection efficiency between spots. It also lets you compare the relative change in gene expression levels, as well as display these levels in a similar scale on the same graph. GeneSpring uses the following formula to normalize to the median for each gene:

$$\frac{\textbf{(signal strength of gene A in sample X)}}{\textbf{(median of every measurement taken for gene A throughout your experiment)}}$$

If the median of the gene's measurements is below the specified cutoff value, the cutoff is used instead. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

The Partially Normalized option means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

GeneSpring does not allow you to perform this normalization and normalize to sample(s), as they address the same issue.

Note: Applying the Normalize to Median technique to data analyzed with the Sample Correlation tool can lead to misleading results in certain circumstances. If the number of replicates in your sample is less than 10, the sample correlation values will be artificially low. This is a side effect of the mathematics used to obtain the normalized expression values, and while it is not wrong, it can lead to misleading results.

If you perform sample correlation analysis on data with a low number of replicates, it is recommended that you not use the Normalize to Median normalization technique. If you do use this technique, ensure that your data includes at least 10 replicates per group.

Median Polishing

Median polishing means that each chip is normalized to its median and each gene is normalized to its median. These normalizations are repeated until the medians converge, up to a maximum of five iterations. This limit prevents endless looping if the normalization coefficients do not converge.

Limiting Measurements by Flag Values

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction.

To limit measurements by flag values:

1. Select the **Use only measurements flagged** box.
2. Select the appropriate option from the menu. The available options are:
    - Present Only
    - Present or Marginal
    - Anything but Absent

Limiting Measurement by a Cutoff Value

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

The Partially Normalized option means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit measurements by a cutoff value:

1. Check the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

Applying Background Correction

You can choose to apply additional background correction in this step.

To apply background correction:

- Select the appropriate box in the Background Correction section of the window. The options are:
  - **Never apply extra background correction**.
  - **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes.
  - **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Normalization Strategies for Specific Technologies

This section describes normalization strategies for the following technologies:

- Normalization of Affymetrix Data
- Normalization of Two-color Microarray Data
- Region Normalization
- Handling Repeated Measurements
- Negative Control Strengths

Normalization of Affymetrix Data

Often data in affymetrix *chp* files are either pre-scaled or are pre-normalized. While Affymetrix's scaling and normalizations are designed to meet the same needs as GeneSpring's, they are not equivalent.

The Affymetrix global scaling procedure, which is comparable to GeneSpring's per-chip normalization, scales the data of each chip to a user-defined target intensity. However, GeneSpring's per chip normalization option, Use Distribution of All Genes, divides each intensity value by the median of all of values on the chip. The resulting expression levels on each chip are centered around 1.

For pre-scaled Affymetrix data, we recommend applying per chip normalizations using the distribution of all genes. In addition we recommend applying a per gene normalization using the median of each gene. The greatest benefit of performing these normalizations is

that each gene intensity is centered artificially around 1. Several GeneSpring functions depend on scaling around 1, especially the Cross-Gene Error Model.

Normalization of Two-color Microarray Data

Like most other technologies, two-color experiment data should be normalized at the gene level (to standardize expression levels between genes), and at the chip level (to standardize expression values between arrays). Two-color experiments are designed to provide an internal standard at the spot level. This per-spot normalization often provides the same scaling that would be provided by a per gene normalization, and thus per-gene normalization is often unnecessary.

Per-chip normalizations are useful in two-color experiments to standardize the global intensities across multiple arrays. Even after applying a per-spot normalization, global variability between chips often remains due to differences in the total amount of the dyes added to the sample and reference samples from one chip to another.

However, the intensity dependent normalization option (which is actually a per gene normalization) succeeds in centering all of the values on each array around 1. In addition, it provides protection from dye incorporation artifacts that lead to unwanted relationships between signal intensity and normalized expression values (see “Intensity-dependent Normalization” on page 10-13).

It is generally recommended to apply a per-spot normalization using the **Divide by control channel** option, followed by selecting the **Per Chip Normalization** step. Both of these normalization options can be accessed by selecting **Experiments > Experiment Normalizations...**.

Region Normalization

The Region normalization option lets you normalize sections of a sample, rather than normalizing over the entire sample. This is especially important if you used multiple arrays for each experimental point or if there is some reason you must normalize sections of an array separately from one another.

When you apply this type of normalization, you normalize over each region, rather than each sample. Hence the formulas for these normalization options become:

Normalizing to Negative Controls for a Region:

$$\frac{\textbf{(the control strength of gene A in region Y of sample X)}}{\textbf{(the median signal of the negative controls in region Y of sample X)}}$$

Normalizing to Positive Controls for a Region:

$$\frac{\textbf{(the control strength of gene A in region Y of sample X)}}{\textbf{(the median signal of the positive controls in region Y of sample X)}}$$

Normalizing Each Region to Itself:

$$\frac{\textbf{(the signal for gene A in region Y of sample X)}}{\textbf{(the median signal of the genes region Y of sample X)}}$$

Handling Repeated Measurements

Occasionally, the raw experimental data in the data file for your sample has more than one line devoted to a particular gene. This may be because you did the sample twice or because you did the sample once but took the measurements twice. If the same gene name is reported multiple times on different horizontal lines in your data file, GeneSpring automatically considers the measurements repeats and averages the signal strengths together.

GeneSpring reports the average and keeps track of the minimum and maximum values for each gene, but it cannot access the particular values falling between the minimum and maximum values.

The formula for averaging a repeated gene is:

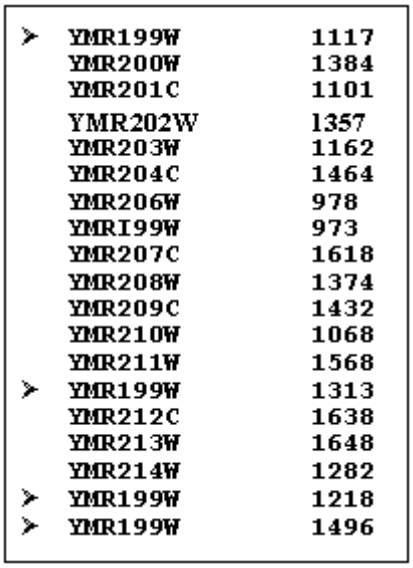
$$\frac{[(\textbf{the signal strength of gene A1}) + (\textbf{the signal strength of gene A2}) + \ldots + (\textbf{the signal strength of gene An})]}{\textbf{N}}$$

This process is repeated for each gene repeated in a data file *before* any other normalizations are applied to the raw values.

Frequently, samples are repeated with exactly the same parameters, but are reported in different data files. If this is the case, the fact that the samples are repeats is represented via parameter. The same normalization is employed when dealing with an experimental parameter considered to be a repeat, but in that case, the averaging takes place after the raw data for each gene has been normalized. See “Setting Up Experiment Parameters” on page 5-37 for more information about repeats reported in separate data files.

Mathematical Illustration

Given this raw data, with four repeats of YMRI99W (marked with the arrows):



➤	YMR199W	1117
	YMR200W	1384
	YMR201C	1101
	YMR202W	1357
	YMR203W	1162
	YMR204C	1464
	YMR206W	978
	YMRI99W	973
	YMR207C	1618
	YMR208W	1374
	YMR209C	1432
	YMR210W	1068
	YMR211W	1568
➤	YMR199W	1313
	YMR212C	1638
	YMR213W	1648
	YMR214W	1282
➤	YMR199W	1218
➤	YMR199W	1496

GeneSpring averages all of the measurements of YMR199W to get a signal strength of 1286. GeneSpring notices the maximum control strength for YMR199W in this sample is 1496 and the minimum is 1117. These values are the end points of YMR199W's error bar which GeneSpring plots when you choose to display error bars in either the graph or the scatter plot displays.

Measurement Flags

Measurement flags are markers in your data set, and data can be assigned as one of four flags:

- Passed (or OK)
- Marginal
- Absent
- Failed
- Unknown

Flags assigned by you when the experiment in entered into GeneSpring:

- **Good Data—**data are present and reliable. Marked with a "P" for passed or "O" for OK.
- **Marginal Data—**data are present, but of unknown or dubious quality. Marked with an "M" for marginal.
- **Absent Data—**There is no data available, and there should have been. Marked with an "A" for absent or "F" for failed.

Flags assigned by GeneSpring:

- **Unavailable Data—**If there is no flag in the column, GeneSpring assigns that measurement a “U”.

Only measurements at the highest available flag level are combined and treated as replicates. The order of flag precedence is **P M U A**. If one or more Ps are present, only Ps are used. If no Ps are present and one or more Ms are present, only Ms are used, etc. Summary statistics are collected over these cases and stored with the corresponding flag. All other flag data are discarded for the gene.

This is done when the experiment is loaded into GeneSpring and is not affected in any way by later choices about which codes are to be used or displayed. The only way to avoid this is to not declare a flag column during data load, in which case the flags are not available for other uses.

For information about measurement flags and how to load them into your experiment, see “Using the Column Editor” on page 5-5.

Negative Control Strengths

Some types of microarray technology report negative control strengths. This is usually the result of subtracting estimated background levels that are larger than the raw signal. This can happen in situations where the expression levels of the gene are low compared to the measurement error. It can also happen when there is background subtraction or when a mismatched probe set has higher intensity levels than the perfect match probe sets.

If negative signal levels occur in a large fraction of the data used for normalization, there can be problems with the normalization, as the median across the normalization set can be very small or even negative. This leads to unreasonable results of normalization. In such cases, which only occur in a few situations, GeneSpring does an extra step in the normalization, where it readjusts the background level for that data by adding a constant to all the raw control strengths in such a way that the 10th percentile of the signal is set equal to 0, before proceeding with the median normalization.

This correction, called the Affine background correction, is applied only when the 10th percentile of the data is more negative than the median of the data is positive. A warning message appears when you first load your data into GeneSpring if this background correction has been applied.

Whether or not the above correction is applied, negative signal levels may still be present for a few measurements. GeneSpring offers the option as the last step of normalization to set these values to zero. Also, when interpreting data in logarithm or fold interpretations, GeneSpring treats all normalized ratio values less than 0.01 (including 0 and negative values) as if they had a ratio of 0.01 to prevent transformation problems.

References

Clevel, W. S., and S. J. Devlin. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association 83*, 596-610.

Yang, Y.H., Dudoit, S., Luu, P., and T.P. Speed. (2001) Normalization for cDNA Microarray Data. *SPIE BiOS 2001*, San Jose, California, January 2001.

11

Annotating Genes

This chapter explains how to work with annotations in GeneSpring. It covers the following topics:

- GeneSpider
- Databases
- Updating Annotations with GeneSpider
- Building Ontologies
- Using Pathways

In addition to the annotation tools described in this chapter, you can also update annotations using the Master Table of Genes Editor. Go to "Managing Genes and Annotations" on page 6-35 for complete information on this editor.

GeneSpider

GeneSpider is a tool that can retrieve annotations using GenBank Accession numbers from three different National Center for Biotechnology Information (NCBI) databases (GenBank, LocusLink, and Unigene). GeneSpider retrieves information about the common name, map position, EC number, description, product, phenotype, function, keywords, sequence (optional), database cross references, and GO annotations, if available, for each gene in a genome. All this information is stored in a standard tabular format in GeneSpring.

These annotations can be used to check each individual gene of interest, without having to go through the literature. They can also be used to classify genes according to keywords present in one or all fields of annotation.

Databases

GeneSpring enables you to update genomic information, including annotations, from the following databases:

- Silicon Genetics
- GenBank
- LocusLink
- UniGene

Silicon Genetics

The Update Annotations from Silicon Genetics function retrieves annotations from the Silicon Genetics Mirror server at info.sigenetics.com. The server downloads the complete databases for GenBank, RefSeq, LocusLink, and UniGene from NCBI. It returns the same annotations as the GeneSpiders that access GenBank, LocusLink, and UniGene, depending on the annotation sources chosen by the user in the Update genome from Silicon Genetics window. In addition, the Silicon Genetics GeneSpider fills in GO biological process, GO molecular function, GO cellular component, and RefSeq identifier fields from the LocusLink database, and the UniGene cluster ID from the UniGene database.

GenBank

The Update Annotations from GenBank function retrieves information from the GenBank and RefSeq databases at NCBI, which both use the same record format. A sample record is shown below, with information retrieved by the GeneSpider. The fields that are filled in the GeneSpring's Master Table of Genes are indicated on the following line. The record is organized by keywords. The record is organized by feature keys.

```
LOCUS HUMELK        1A   2266 bp    mRNA    linear   PRI 07-NOV-1994

DEFINITION          Homo sapiens tyrosine kinase (ELK1) oncogene mRNA
                    complete cds

                    ^ Description field in GeneSpring

ACCESSION           M25269

VERSION             M25269.1 GI:538208

KEYWORDS            ETS1 gene; oncogene; tyrosine kinase

                    ^ Keywords field in GeneSpring

SOURCE              Homo sapiens cDNA to mRNA

ORGANISM            Homo sapiens

                    Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
                    Euteleostomi; Mammalia; Eutheria; Primates;
                    Catarrhini; Hominidae; Homo

REFERENCE           1 (bases 1 to 2266)

AUTHORS             Rao,V.N., Huebner,K., Isobe,M., ar-Rushdi,A.,
                    Croce,C.M. and Reddy,E.S.

TITLE               elk, tissue-specific ets-related genes on
                    chromosomes X and 14 near translocation breakpoints

JOURNAL             Science 244 (4900), 66-70 (1989)

MEDLINE             89203250

PUBMED              2539641
```

COMMENT	On Sep 15, 1994 this sequence version replaced gi:341319
FEATURES	Location/Qualifiers
source	1..2266 /organism=”Homo sapiens” /db_xref=”taxon:9606” /map=”Xp22.1-p11” **^ Map field in GeneSpring** /clone=”lambda-11” /cell_line=”COLO 320”
gene	1..2266 /gene=”ELK1” **^ Common field in GeneSpring**
CDS	316..1602 /gene=”ELK1” **^ Common field in GeneSpring** /codon_start=1 /product=”tyrosine kinase” **^ Product field in GeneSpring** /protein_id=”AAA52384.1” /db_xref=”GI:538209” **^ DBId field in GeneSpring** /db_xref=”GDB:G00-119-867” **^ DBId field in GeneSpring** // translation=”DPSVTLWQFLLQLLREQGNGHIISWTSRDGGEFKLVDA EEVAR LWGLRKNKTNMNYDKLSRALRYYYDKNIIRKVSGQKFVYKFVSYPEVAGCS TEDCPPQ PEVSVTSTMPNVAPAAIHAAPGDTVSGKPGTPKGAGMAGPGGLARSSRNEY MRSGLYS TFTIQSLQPQPPPHPRPAVVLPNAAPAGAAAPPSGSRSTSPSPLEACLEAE EAGLPLQ VILTPPEAPNLKSEELNVEPGLGRALPPEVKVEGPKEELEVAGERGFVPET TKAEPEV PPQEGVPARLPAVVMDTAGQAGGHAASSPEISQPQKGRKPRDLELPLSPSL LGGPGPE RTPGSGSGSGLQAPGPALTPSLLPTHTLTPVLLTPSSLPPSIHFWSTLSPI APRSPAKLSFQFPSSGSAQVHIPSISVDGLSTPVVLSPGPQKP”
ORIGIN	1 aattccgagc tgtagggaaa cgcaggggcg gcttctaggt gctgccgccg ccaccgccac **^ Sequence field in GeneSpring** [remaining sequence omitted]

The GeneSpider fills in annotations from GenBank as summarized in the record. Feature keys are found in the margin of the record (the section between the FEATURES line and the ORIGIN or BASE COUNT line). Qualifiers indicate information about a feature; they begin with a slash followed by the qualifier name, then an equals sign (for example, `/gene=`).

A complete description of the format for the latest release of GenBank is available at ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt.

LocusLink

Update Annotations from LocusLink reads the HTML source from queries to the LocusLink database at NCBI as summarized in Table 11-1.

Table 11-1 LocusLink Sample Record Format

Annotation	LocusLink Label
Common	Official Gene Symbol or Interim Gene Symbol plus any Alternate Symbols
Map	Position, Cytogenetic, or Chromosome
EC number	EC number
Product	Product
Phenotype	Phenotype

UniGene

Update Annotations from UniGene reads the html source from queries to the UniGene database at NCBI. The Common Name and Description are from the line immediately following the UniGene cluster ID and the species name. If only one item is shown, it is returned as the description. The map annotation is taken from the Cytogenetic Position under the MAPPING INFORMATION heading; if Cytogenetic Position is not given, the Chromosome number is used.

Updating Annotations with GeneSpider

After you have loaded a new genome, you can make sure that it contains the latest information from the genome databases on the World Wide Web by using GeneSpider. To use GeneSpider, you must have GenBank accession numbers in your Master Table of Genes. GenBank accession numbers are usually added to the GenBank column of the master gene table. If you have multiple GenBank accession numbers for a single gene, they should be separated by semicolons.

For more information about the Master Table of Genes, go to “Managing Genes and Annotations” on page 6-35.

Limitations on Using GeneSpiders

NCBI has the following requirement for automated access:

- Run retrieval scripts on weekends or between 9 PM and 5 AM ET weekdays for any series of more than 100 requests.

Refer to the NCBI disclaimer and copyright notice at http://www.ncbi.nlm.nih.gov/About/disclaimer.html for more information. See http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html for more information.

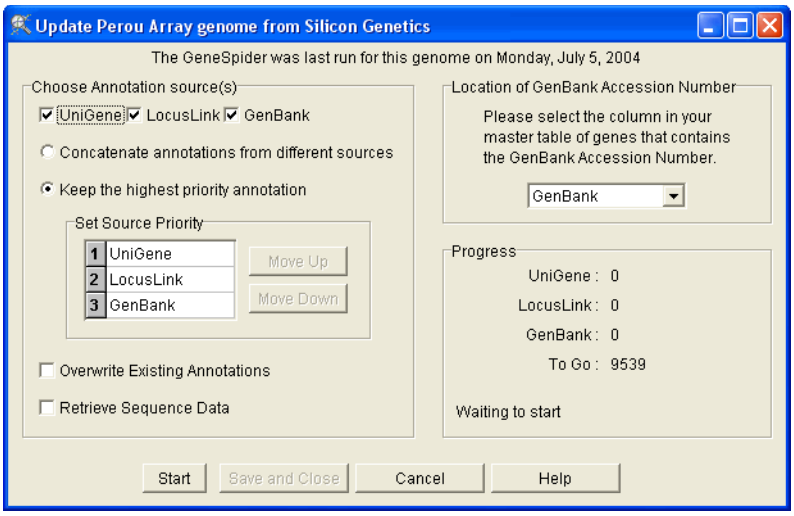
The Silicon Genetics mirror server can be used any time of the day.

Updating Annotations from Silicon Genetics

To update annotations from Silicon Genetics:

1. Select **Annotations > GeneSpider > Update Annotations from Silicon Genetics**.

   The Update Genome from Silicon Genetics window opens.



   The date when GeneSpider was last run for this genome is indicated at the top of the window.

2. In the **Choose Annotation Source** box, select the source for the annotations.
3. In the **Location of GenBank Accession Number** box, select the column in your Master Table of Genes that contains the accession number.

   Note: GeneSpring will attempt to guess the best column for the GenBank Accession Number. If you don't like the guess, select the column that you want.

4. In the **Choose Annotations Source** box, do one of the following:
   - Select **Concatenate annotations from different sources** to retrieve annotations from all sources as a semicolon-delimited list. Exact duplicates are not retrieved. The order is fixed: GenBank, then LocusLink, then UniGene.
   - Select **Keep the highest priority annotation** to retrieve only the annotation from the highest priority source available for each gene.
5. To update information in places where data already exists, select the **Overwrite Existing Annotations** check box.

   If you leave this box unchecked, GeneSpring adds new information only to blank fields.
6. Click the **Start** button to begin updating annotations.

   While the GeneSpider runs there are a number of informational fields visible.
   - **GenBank, UniGene, LocusLink**—Number of genes whose annotations have been retrieved from the given source database.
   - **Processed—**Number of genes in the genome that the GeneSpider has finished querying the database on.
   - **Found**—number of processed genes where the GeneSpider has found a useful record in the database.
   - **Enhanced—**Number of genes where information has been found and added to the Master Table of Genes.
   - **To Go—**Number of genes in the genome that have not been processed.
7. Click the **Save and Close** button.

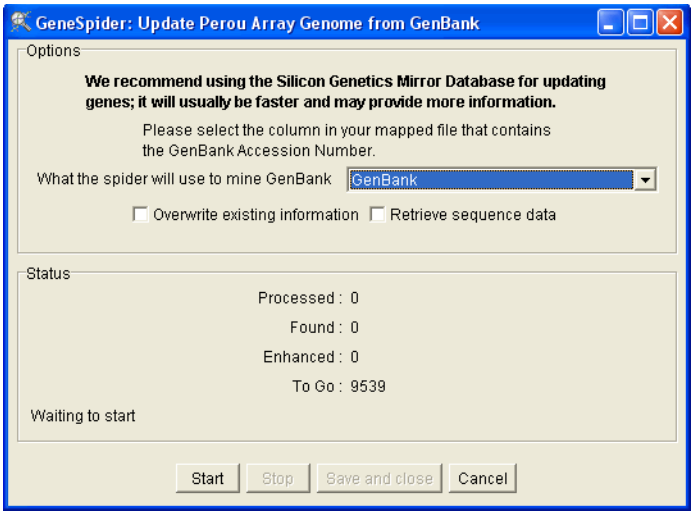
   The Master Table Of Genes is not permanently updated until you click **Save and Close**. This button is inactive while the GeneSpider is running. You must wait until the GeneSpider is finished, before clicking **Save and Close**. When you save annotations, GeneSpring maintains a backup file of the old Master Table of Genes.

Updating Annotations from GenBank, LocusLink, or UniGene

To update annotations from GenBank, LocusLink, or UniGene:

1. Select **Annotations > GeneSpider**.
2. Do one of the following:
   - Select **Update Annotations from GenBank**.
   - Select **Update Annotations from LocusLink**.
   - Select **Update Annotations from UniGene**.

The GeneSpider: Update Genome window opens.


GeneSpider: Update Perou Array Genome from GenBank
Options
We recommend using the Silicon Genetics Mirror Database for updating genes; it will usually be faster and may provide more information.
Please select the column in your mapped file that contains the GenBank Accession Number.
What the spider will use to mine GenBank
GenBank
Overwrite existing information
Retrieve sequence data
Status
Processed : 0
Found : 0
Enhanced : 0
To Go : 9539
Waiting to start
Start
Stop
Save and close
Cancel

3. In the **Options** box, select the annotation source you want from the **What the spider will use to mine the source** menu.

   Note: GeneSpring will attempt to guess the best column for the GenBank Accession Number. If you don't like the guess, select the column that you want.

4. To update information in places where data already exists, select the **Overwrite existing information** check box.

   If you leave this box unchecked, GeneSpring adds new information only to blank fields.

5. To update genomic sequences, select the **Retrieve Sequence Data** option.

6. Click the **Start** button to begin updating annotations.

   While the GeneSpider runs there are a number of informational fields visible.

   - **Status—**this is the level of completion the GeneSpider has reached.
   - **Processed—**Number of genes in the genome that the GeneSpider has finished querying the database on.
   - **Found**—number of processed genes where the GeneSpider has found a useful record in the database.
   - **Enhanced—**Number of genes where information has been found and added to the Master Table of Genes.
   - **To Go—**Number of genes in the genome that have not been processed.

7. Click the **Save and Close** button.

   The Master Table Of Genes is not permanently updated until you click **Save and Close**. This button is inactive while the GeneSpider is running. You must wait until the GeneSpider is finished, before clicking **Save and Close**. When you save annotations, GeneSpring maintains a backup file of the old Master Table of Genes.

GeneSpider Errors Window

While the GeneSpider is running, the GeneSpider Errors window may appear. This window lists any errors the GeneSpider encountered and a brief description of the problem. For example, if no match was found for some genes on your system, the Errors window displays the gene identifier and the text “Gene not found”.

The most common reason for no genes to be updated is that you did not select the annotation column containing the GenBank Accession Numbers.

Problems with the Map Location Annotations

This window opens when the information the spider has retrieved for a gene’s map location field does not meet the required criteria. GeneSpring offers suggested corrections in the Corrected Map Location column of the table. You can edit the entries in this column, if necessary.

Which Annotations are Retrieved?

Table 11-2 describes the annotation fields retrieved by GeneSpider from the various databases:

Table 11-2 NCBI Database Annotation Fields

Annotation	Silicon Genetics	GenBank	LocusLink	UniGene
Systematic name*				
Common	X	X	X	X
Map	X	X	X	X
GenBank*				
Synonym*				
EC number	X	X	X	
Description	X	X		X
Product	X	X	X	
Phenotype	X	X	X	
Function	X	X		
Keywords	X	X		
PubMedID*				
Type**				
DBId	X	X		
GO biological process †	X			
GO molecular function †	X			
GO cellular component †	X			
RefSeq	X			

Table 11-2 NCBI Database Annotation Fields (Continued)

UniGene***	X			
Sequence	X	X		

* Systematic name, GenBank accession number, Synonym, and PubMedID are not filled in by GeneSpider. These fields can be filled in manually when a genome is created.

** Type is filled in by GeneSpring when it reads a genome from a GenBank (gbk) file. The value is commonly “CDS”, but “mRNA”, “rRNA”, “terminator”, “gene”, and other GenBank feature keynames are possible entries.

*** UniGene is also filled in by the Build Homology Tables feature. When the user requests Build Homology Tables to save UniGene cluster IDs, it replaces the UniGene column with its results, deleting any obsolete entries from the column. (This contrasts with the behavior of the GeneSpider, where overwrite will not replace existing entries with an empty entry.)

† Gene Ontology information on the Silicon Genetics Mirror server is obtained from the LocusLink database. The same information cannot be provided through Update Annotations from LocusLink because the LocusLink web page format does not indicate the organizing principles of biological process, molecular function, and cellular component.

Building Ontologies

GeneSpring leverages data found in publicly available genomics databases to build gene ontologies based on the latest annotation information available. These ontologies provide insights about possible cellular activity associated with your expression data.

The Build Simplified Ontology tool hierarchically groups genes into meaningful biological categories (gene lists), based on the SLIMS version of GO. To form these groups, GeneSpring’s ontology tool parses all of the annotations in the genome. It then assigns each gene to one or more ontology groups based on this analysis.

The Build Simplified Ontology constructor builds over 300 biologically meaningful gene lists that can be compared, merged or browsed. These lists can be further used to annotate clusters and cross-reference new gene lists.

To build an ontology:

1. Select **Annotations > Build Ontology**.
2. Enter a name for the new ontology folder, or leave the default name to overwrite the existing simplified ontology list.
3. Click the **OK** button.

The new ontology list appears in the Gene Ontology (GO SLIMS) subfolder under the Gene Lists folder.

Using Pathways

With the Pathway viewer, genes and their expression patterns can be visually characterized based on their location within a cellular pathway. You can design you own pathway diagrams or directly import publicly available pathway maps. You can predict genes associated with discrete steps in the pathway of interest.

GeneSpring can use graphical representations of pathways to map genes and their expression values. The pathway is usually imported as a picture file. Genes are then manually mapped onto the pathway image and can be viewed in the context of your microarray experiment. This can be particularly useful if you are interested in genes whose differential expression directly affects the regulation of other genes on the pathway.

When using pathways, genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context. You can zoom in on a pathway and move the slider to watch gene expression change over the experimental conditions. One scenario in which a pathway can be very useful is if you are trying to identify a class of genes that are associated with a particular step or regulatory element within a pathway (Figure 11-1).

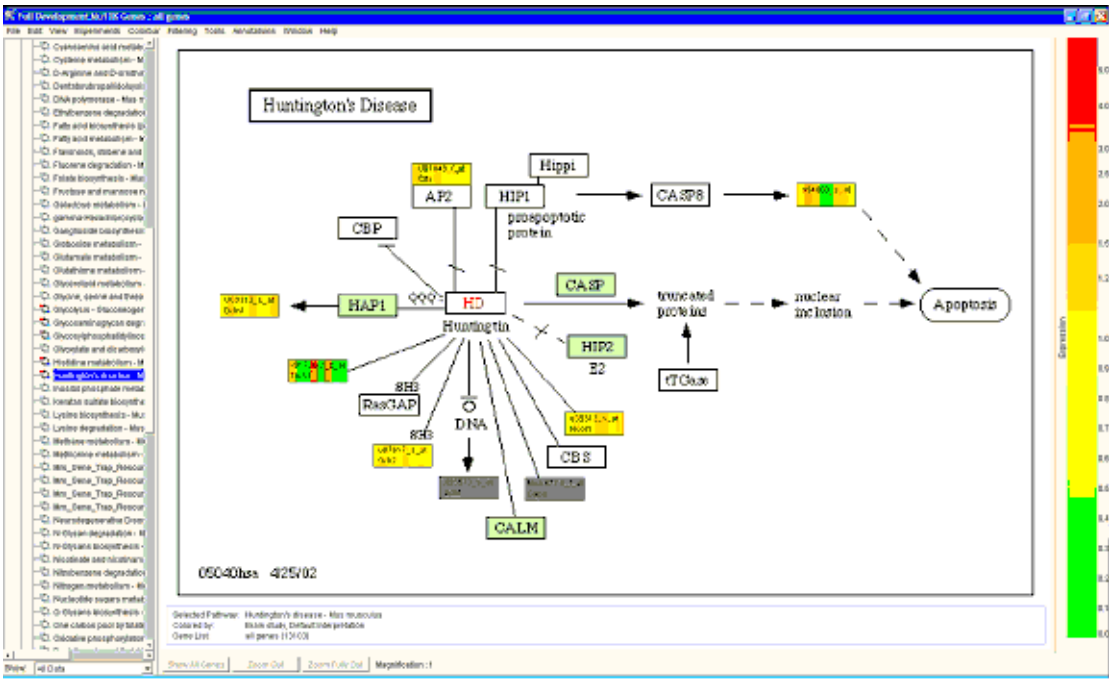


Figure 11-1 Cell Cycle Pathway

The pathway view also enables you to find genes with a similar expression pattern. Finally, you can display gene expression across all conditions, to view how differentially regulated genes in certain conditions might affect downstream genes.

Gene Expression data analyzed in GeneSpring can be visualized on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways or GenMAPP pathways.

KEGG Pathways

KEGG is a large public database containing pathways for many organisms, all of which are available for download via FTP. To locate pathways for a specific organism, go to the following URL: ftp://ftp.genome.ad.jp/pub/kegg/pathways.

Folders are named using three letter abbreviations based on the Latin name of the organism, for example, “hsa” is Homo sapiens, “mmu” is Mus musculus.

To locate generic pathways, download the “maps” folder. This folder contains reference pathways (metabolic and regulatory). These pathways contain enzymes rather than genes. When you import an organism-specific pathway, the GENE file is parsed, and for each KEGG accession number the gene names are stored (including the accession number itself, if it is not fully numeric). All EC numbers are also stored.

GeneSpring first tries to match the first gene name with the genes in the current genome. It is compared against the Systematic Names and Common Names. If no match is found, GeneSpring attempts to match the second name, and so on. As soon as matching genes are found (often it is one gene, but sometimes there are more), GeneSpring stops any further processing in order to decrease the chance of false matches.

If no matches are found, GeneSpring tries to match EC numbers. Because several genes may correspond to the same EC number, in this case GeneSpring searches for matches to all EC numbers in the GENE file. This reduces the possibility of a false match.

GenMAPP Pathways

Silicon Genetics has reached an agreement with the GenMAPP organization to redistribute the pathways for use in GeneSpring. Genes will be automatically mapped to the pathways based on their gene symbol.

GeneSpring-formatted pathways can be downloaded from http://www.silicongenetics.com/cgi/SiG.cgi/Support/resources.smf. For more information on the GenMAPP organization, go to http://www.genmapp.org.

Importing KEGG Pathways

To import a KEGG pathway:

1. Select **File > New Pathway > Import KEGG Pathway**.

   The Import KEGG Pathways window opens.

2. Navigate to the directory that contains the pathways you downloaded from KEGG.
3. (Optional) Select the check box to group pathways alphabetically into subfolders of nine pathways each.

   This step is useful if you want to split a window by a group of pathways.

4. (Optional) Select the check box to create gene lists from the imported pathways.

   If you checked the box to group pathways into subfolders, the gene lists will also be grouped by subfolders.

5. Click **OK**.

6. In the window that appears, accept the default or enter a new name for the folder of pathways.
7. Click **OK**.

Saving pathways may take several minutes.

Importing JPEG and GIF Pathway Image Files

To import a JPEG or GIF pathway image file:

1. Select **File > New Pathway> Import New Pathway**.
2. Browse for the pathway image.

Genes will not be mapped onto pathways automatically, and should be added to the image manually.

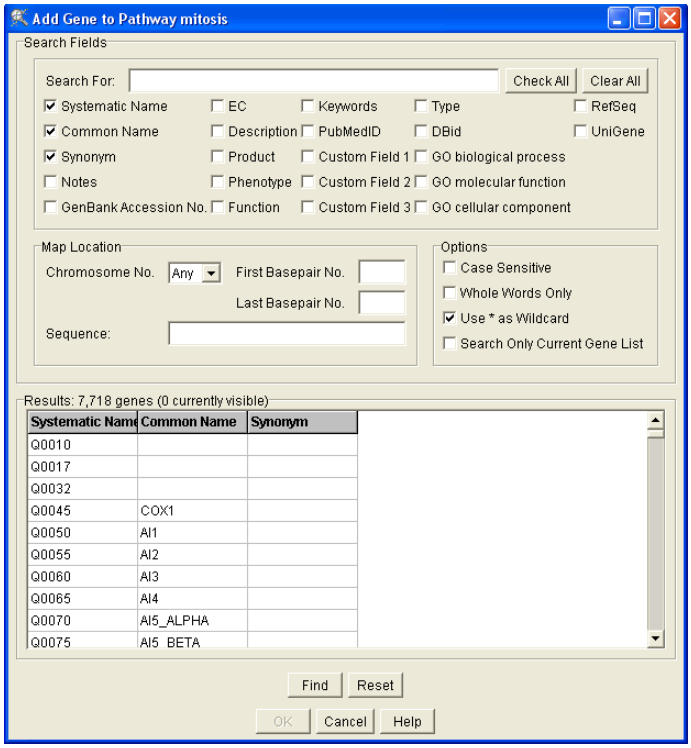
Adding Genes to Pathways

Once you have successfully imported your pathway into GeneSpring, you can place genes on top of the background image.

To add genes to a pathway:

1. In the **Navigator**, open the **Pathways** folder, and then select the pathway you want.
2. Do one of the following:
   - (Windows) Hold down **Ctrl** and draw a rectangle where you would like the gene to appear on the pathway.
   - (Macintosh) Hold down Option-click and draw a rectangle where you would like the gene to appear on the pathway.

The Add Gene to Pathway window opens.



3. In the **Annotation** check boxes, do one of the following:
   - To search individual annotations, select the check box you want.
   - To search all annotations, click the **Check All** button.
   - To clear all selected annotations, click the **Clear All** button.
4. In the **Search For** field, enter the search terms you want:
   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.
   - Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified.
   - You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.

- To perform an exact string match, use double quotes in your search term. For example, “expression profile”.
- For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

5. To restrict your search, you can select any of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
   - **Search Only Current Gene List**—Searches the current gene list only.
6. Click the **Find** button.

   A list of the genes that match the search criteria is displayed in a table at the bottom of the window.
7. Select the gene you want in the search results and click the **OK** button.

   The gene name appears on the pathway.
8. If the gene name or keyword is present for more than one gene, another window opens directing you to choose a gene ID from a list. Double-click on the appropriate ID.
9. To remove a gene, right-click on the element and select the **Delete Pathway Element** command.

Making Gene Lists from Pathways

GeneSpring uses proprietary algorithms to predict the genes that fit near a selected point on a pathway. When you select a point, GeneSpring makes two lists of genes from those currently displayed on your diagram. List A contains the two genes that appear closest to your selected point on the diagram. List B contains all other genes on the pathway.

GeneSpring examines all the genes on your currently selected gene list and finds all genes whose minimum similarity (correlation) with genes on list A is higher than their maximum similarity with genes on list B. These genes are made into a separate gene list for you to examine. You can place a gene from this list on the pathway (see "Adding Genes to Pathways" on page 11-12).

If your pathway geometry is complex, this procedure is not very useful, since it relies on window distance only, not pathway structure or connectivity.

To find new genes on a pathway:

1. In the **Navigator**, open the **Pathways** folder, and then select the pathway you want.
2. Right-click near a group of genes displayed on your pathway.
3. Choose the option **Find Genes Which Could Fit Here**.

   The New Gene List window opens.
4. Enter a name and destination folder and click the **Save** button.

   The new gene list is saved in your Gene Lists folder.

Pathway Commands

Right-click your Pathway in the Navigator for the following options:

- **Display Pathway—**Displays the selected pathway in the Genome Browser.
- **Make Gene List—**lets you save a list of all the genes on the selected pathway.
- **Attachments—**lets you add a text or picture attachment to your Pathway.
- **Inspect—**Displays a listing of details such as pathway history and genome.
- **Upload to Signet—**Uploads your information and the pathway picture to Signet (see "Uploading Data to Signet" on page 18-5.
- **Assign Project**—Assigns the pathway to a project.
- **Rename Pathway—**lets you rename your pathway.
- **Delete Pathway—**Deletes a pathway. A confirmation dialog box appears.

  Export as Zip—lets you export the pathway as a GeneSpring zip file.

12

Working with Gene Lists

This chapter explains how to use gene lists to mine data for further analysis. It covers the following topics:

- Gene List
- Managing Gene Lists
- Filtering Gene Lists
- Inspecting Gene Lists
- Making Gene Lists
- Working with Homology Tables

Gene List

A gene list is a list of gene names or identifiers created and saved in GeneSpring. Gene lists appear in the Gene List folder of the Navigator. In GeneSpring, gene lists can be used to:

- Reduce data to manageable size
- Find interesting genes
- Filter out bad data (low or high intensity values or abnormal spots on an array)
- Make Venn diagrams by combining different gene lists
- Cluster and characterize data
- Perform statistical analyses
- Mine data with other filtering tools

Managing Gene Lists

This section explains how to use the Gene List Editor window to manage gene lists. It covers the following topics:

- Displaying Gene Lists
- Gene List Editor Window
- Opening the Gene List Editor
- Creating or Editing Gene Lists
- Saving Gene Lists
- Deleting Gene Lists

Displaying Gene Lists

You can display gene lists using the Navigator in the main GeneSpring window.

Displaying a Gene List

To display a gene list:

- Do one of the following:
  - In the **Navigator**, open the **Gene Lists** folder, and then select the gene list you want.
  - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list, and then select **Display**.

Displaying a Gene List as a Secondary List

The **Display as Second List** offers a convenient way to compare two gene lists quickly. When a secondary gene list is displayed, genes that are in the secondary list, but not in the primary gene list, are colored by the Structure Color settings defined in your preferences. Go to "Setting Preferences" on page 3-15 for more information

To display a gene list as a secondary list:

- In the **Navigator**, open the **Gene Lists** folder, right-click the gene list, and then select **Display as Second List**.

To remove the secondary gene list:

- Select **View** > **Remove Secondary Gene List**.

Showing All Genes

At any time in any display mode, you can click the **Show All Genes** button, located at the bottom of the Genome Browser, to revert to a display of all genomic elements.

Gene List Editor Window

You manage the contents of gene lists using the Gene List Editor window (Figure 12-1). From this window you can create a new gene list based on a variety of selection criteria, add genes to an existing gene lists, or remove genes from gene lists.

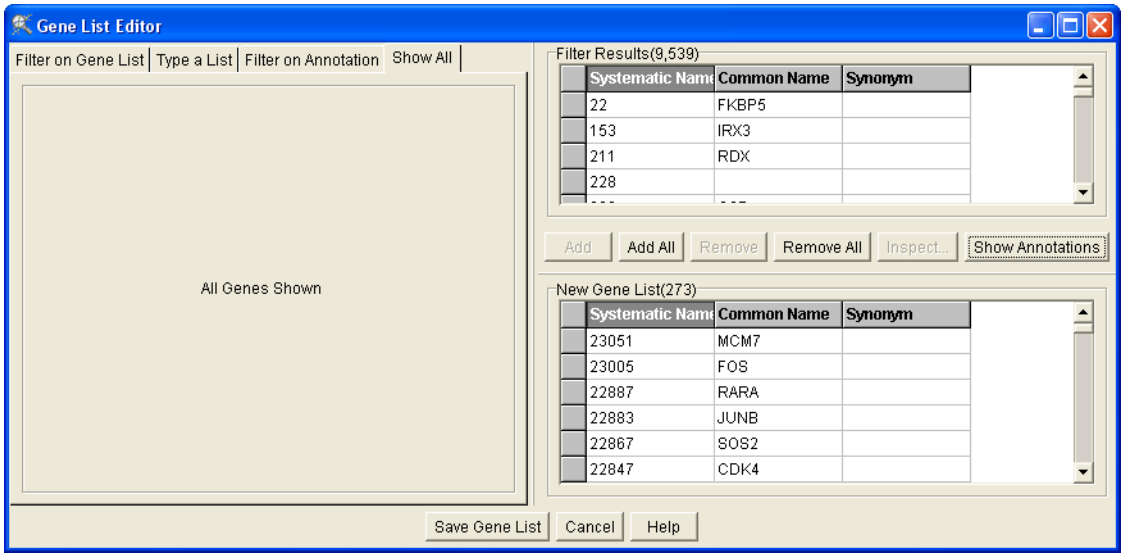


Figure 12-1 Gene List Editor Window

Filter Options

The left side of the window contains tabs for each filtering method. Click a tab to view options for that method. These methods are:

- **Filter on Gene List**—Display genes from a selected gene list.
- **Type a List**—Manually enter a list of genes.
- **Filter on Annotation**—Display genes based on a specified annotation.
- **Show All**—Display all available genes without applying a filter.

Filter Results Table and New Gene List Table

The right section of the window contains two tables: Filter Results and New Gene List. The upper table contains all of the genes resulting from the current filtering method. The lower table contains the genes you have selected to add to your list. Between the two tables are six buttons.

Buttons

- **Add**—Adds a selected gene in the Filter Results table to the New Gene List table.
- **Add All**—Adds all genes in the Filter Results table to the New Gene List table.
- **Remove**—Removes a selected gene from the New Gene List table.
- **Remove All**—Removes all genes from the New Gene List table.

- **Inspect**—Displays the selected gene in the Gene List Inspector. Go to "Using the Gene List Inspector" on page 7-33 for more information.
- **Show/Hide Annotations**—Shows or hides annotations from a gene list.

Filter on Gene List Tab

To filter on gene list, use the Filter on Gene List tab (Figure 12-2) to locate the desired gene list and select it in the list. All genes in that list appear in the Filter Results table.

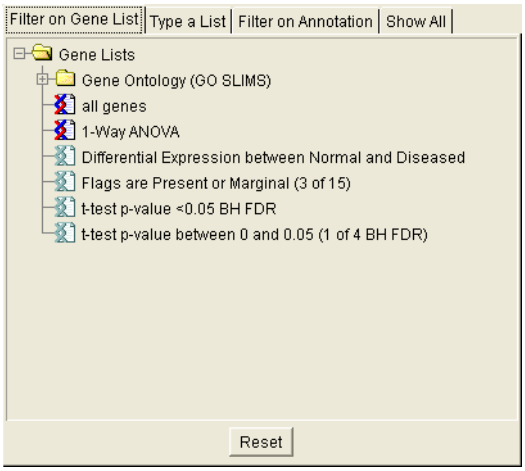

Filter on Gene List
Type a List
Filter on Annotation
Show All
Gene Lists
Gene Ontology (GO SLIMS)
all genes
1-Way ANOVA
Differential Expression between Normal and Diseased
Flags are Present or Marginal (3 of 15)
t-test p-value <0.05 BH FDR
t-test p-value between 0 and 0.05 (1 of 4 BH FDR)
Reset

Figure 12-2 Gene List Editor Window: Filter on Gene List Tab

Type a List Tab

The Type a List tab (Figure 12-3) allows you to manually enter a list of genes. To enter genes, simply click in the gene list box, type a gene's Common Name, Systematic Name, Synonym, or GenBank Accession Number and press **Enter**. You can also use copy and paste to enter one or more genes to this list.

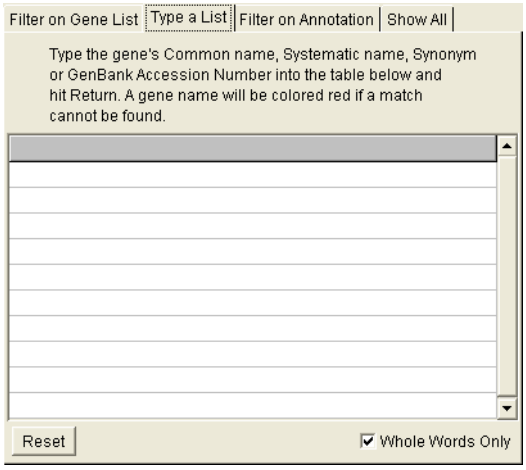

Filter on Gene List
Type a List
Filter on Annotation
Show All
Type the gene's Common name, Systematic name, Synonym or GenBank Accession Number into the table below and hit Return. A gene name will be colored red if a match cannot be found.
Reset
Whole Words Only

Figure 12-3 Gene List Editor Window: Type a List Tab

If the gene is found, it appears in the Filter Results table. If the gene is not found, it is colored in red. To clear your entries, click the **Reset** button. This removes all entered genes from both the list of genes you have typed in and the Filter Results table.

Filter on Annotation Tab

The Filter on Annotation tab (Figure 12-4) allows you to filter genes based on text or values in a specified annotation.

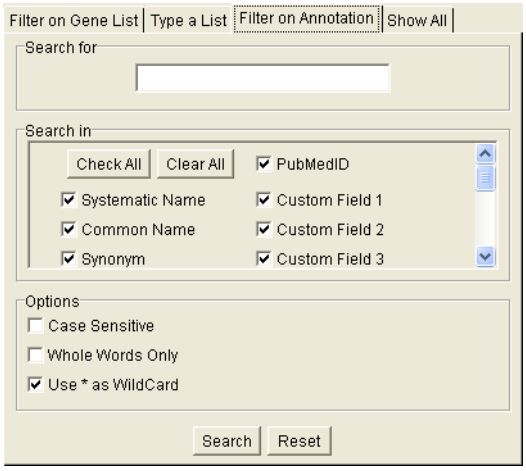


Figure 12-4 Gene List Editor Window: Filter on Annotation Tab

Opening the Gene List Editor

During analysis, you will create and work with interesting collections of genes known as gene lists. These gene lists are stored in the Gene Lists folder. By default, GeneSpring makes and displays an “all genes” list containing all genes in the genome.

To open the Gene List Editor:

- Do one of the following:
  - In the **Navigator**, open the **Gene List** folder, and then select the gene list you want.
  - Select **Edit > Edit Gene List**.

Creating or Editing Gene Lists

This section explains how to create or edit gene lists using the Gene List Editor window. To use other function for making gene lists, go to “Making Gene Lists” on page 12-16 for more information.

To create or edit a gene list:

1. Select **Edit > Edit Gene List**.

   The Gene List Editor window opens.

2. In the **Navigator**, select the gene list you want.

3. To locate the genes want to add to the gene list, select a filter method to use.

   Go to "Filtering Gene Lists" on page 12-8 for information.

   Genes matching the specified search parameters appear in the Filter Results table.

4. In the **Filter Results** table, do one of the following:

   - To add individual genes to the gene list, select the genes you want and click the **Add** button.
   - To add all genes to the gene list, click the **Add All** button.
   - To remove individual genes from the list, select the genes you want to remove and click the **Remove** button.
   - To remove all genes added to the list, click the **Remove All** button.

5. To save the gene list, go to "Saving Gene Lists" on page 12-6.

Saving Gene Lists

The Save Gene List window (Figure 12-5) lets you save a gene list after you have created or edited one.

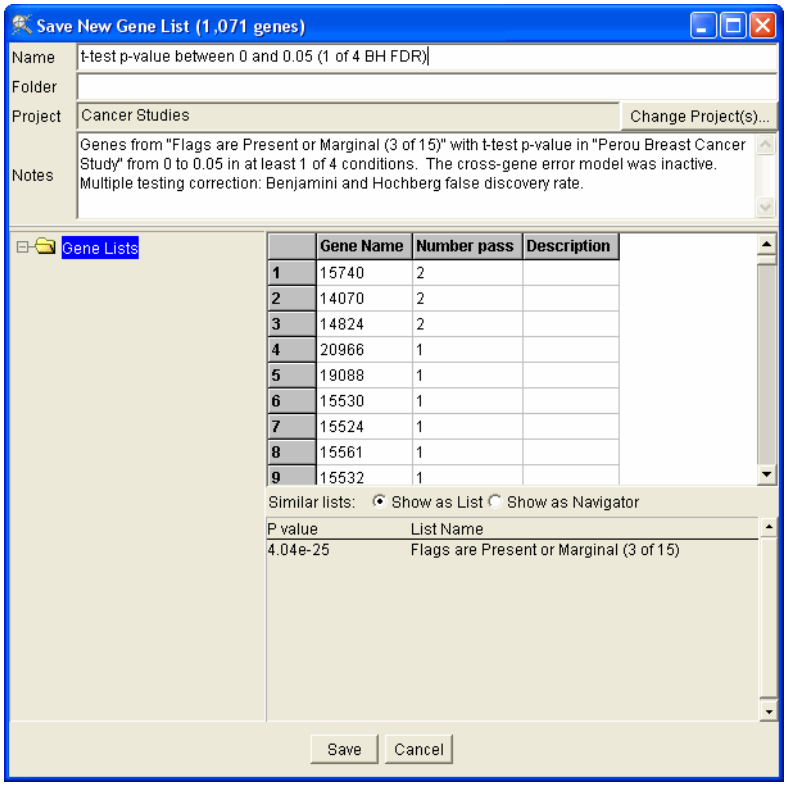

Save New Gene List (1,071 genes)
Name
t-test p-value between 0 and 0.05 (1 of 4 BH FDR)
Folder
Project
Cancer Studies
Change Project(s)...
Notes
Genes from "Flags are Present or Marginal (3 of 15)" with t-test p-value in "Perou Breast Cancer Study" from 0 to 0.05 in at least 1 of 4 conditions. The cross-gene error model was inactive. Multiple testing correction: Benjamini and Hochberg false discovery rate.
Gene Lists
Gene Name
Number pass
Description
1 15740 2
2 14070 2
3 14824 2
4 20966 1
5 19088 1
6 15530 1
7 15524 1
8 15561 1
9 15532 1
Similar lists:
Show as List
Show as Navigator
P value
List Name
4.04e-25
Flags are Present or Marginal (3 of 15)
Save
Cancel

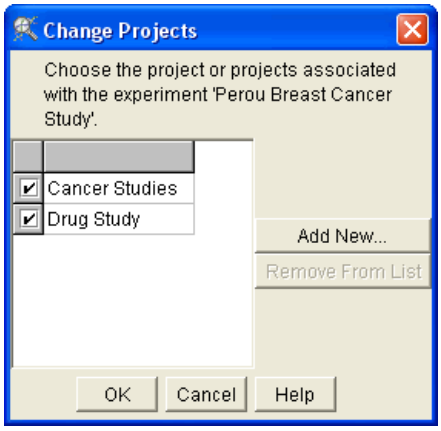
Figure 12-5 Save Gene List Window

To save a gene list:

1. Do one of the following:
   - In the **Navigator**, open the **Gene List** folder, and select the gene list you want.
   - Select **Edit > Edit Gene List**.
2. In the **Name** field, enter a name for the gene list.

   Be sure to choose a descriptive name that you will remember later.
3. To save the gene list in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.
4. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.
5. Enter **Notes** containing more descriptive information about the gene list.
6. To assign the gene list to a project, click the **Change Projects** button.

   The Change Projects window opens.



7. Do one of the following:
   - To create a new project name, do the following:
     a. Click the **Add New** button.

        The Add New window opens.
     b. In the **Project Name** field, enter a name for the project.

        All data objects in the project will be associated with this project name.
     c. Click the **OK** button.
   - To assign a previously created project name, do the following:
     a. Select the project name(s) you want.

        You can assign data objects to more than one project.
     b. Click the **OK** button.
8. Click the **Save** button.

Deleting Gene Lists

To delete a gene list:

1. In the **Navigator**, open the **Gene List** folder.
2. Right-click the gene list you want to delete and select the **Delete** command.

Filtering Gene Lists

This section explains how to use different filters to locate genes you want to include in a gene list. The options are:

- Filtering on Gene Lists
- Filtering on Lists
- Filtering on Annotations
- Showing All Genes

Filtering on Gene Lists

To filter on a gene list:

1. Do one of the following:
   - In the **Navigator**, open the **Gene List** folder, and then select the gene list you want.
   - Select **Edit > Edit Gene List**.

     The Gene List Editor opens.
2. Click the **Filter on Gene List** tab.
3. In the **Gene Lists** folder, select the list you want.

   Genes matching the specified search parameters appear in the Filter Results table.

Filtering on Lists

To filter on a list:

1. Do one of the following:
   - In the **Navigator**, open the **Gene List** folder, and then select the gene list you want.
   - Select **Edit > Edit Gene List**.

     The Gene List Editor opens.
2. Click the **Type a List** tab.
3. Enter the gene's Systematic Name, Common Name, Synonym, or GenBank ID.
4. To search by whole word, click the **OK** button.

   Genes matching the specified search parameters appear in the Filter Results table.

Filtering on Annotations

To filter on annotations:

1. Do one of the following:
   - In the **Navigator**, open the **Gene List** folder, and then select the gene list you want.
   - Select **Edit > Edit Gene List**.

     The Gene List Editor opens.
2. Click the **Filter on Annotations** tab.
3. In the **Search For** field, enter the search terms you want:
   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.
   - Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified.
   - You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
   - To perform an exact string match, use double quotes in your search term. For example, “expression profile”.
   - For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
   - You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.
4. Select the annotation fields to search.

   Some fields may not be visible due to window size. To view all fields, scroll down in the **Search For** text box. To check all boxes, select the **Check All** option. To clear all boxes, select the **Clear All** option.

5. To restrict your search, you can select any of the following options:
   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
6. Click the **Search** button.

   Genes matching the specified search parameters appear in the Filter Results table.

Showing All Genes

To show all genes:

1. Do one of the following:
   - In the **Navigator**, open the **Gene List** folder, and then select the gene list you want.
   - Select **Edit > Edit Gene List**.

     The Gene List Editor opens.
2. Click the **Show All** tab.

   Genes matching the specified search parameters appear in the Filter Results table.

Inspecting Gene Lists

You can view the contents of a gene list, the gene's annotations, and the method used to create the gene list with the Gene List Inspector window.

Gene List Inspector Window

The Gene List Inspector window (Figure 12-6) displays descriptive information about the selected gene list.

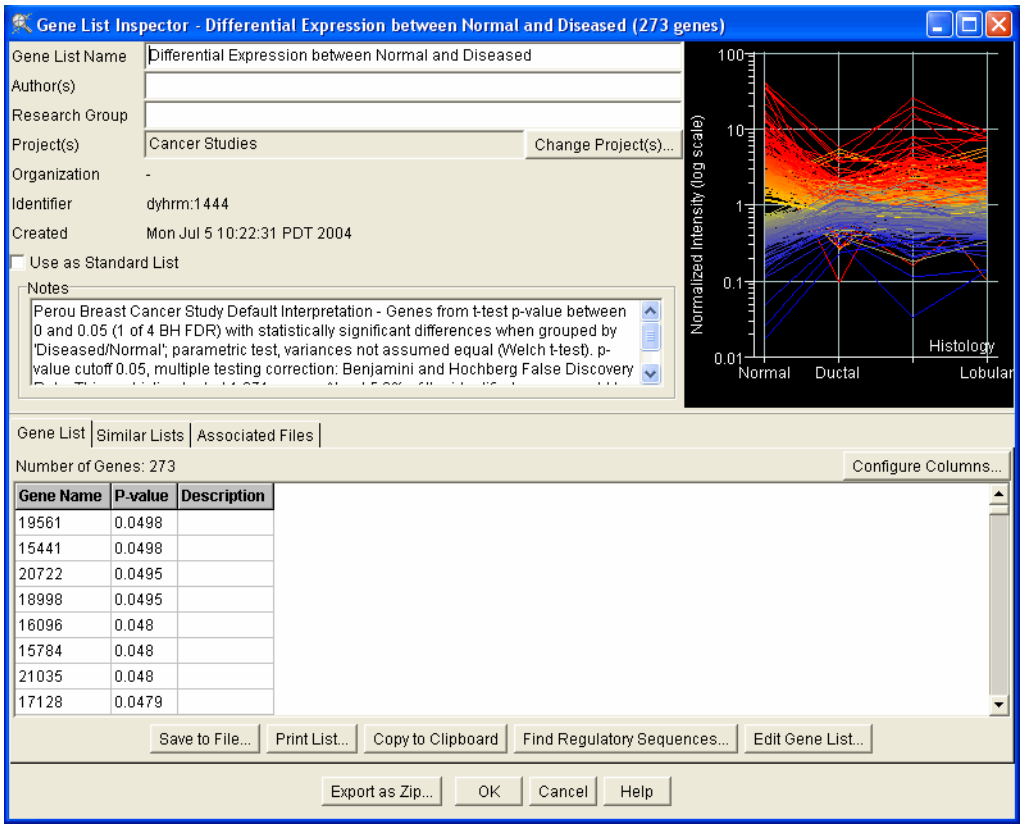


Figure 12-6 Gene List Inspector Window

The following section describes the elements in the Gene List Inspector window.

Summary Information Section

At the top of the Gene List Inspector window are fields that enable you to add information about the genome.

- **Name**—Show the name of the gene list.
- **Author(s)**—Lets you enter the name of the authors associated with the gene list.
- **Research Group**—Lets you enter the name of the research group associated with the gene list.

- **Project—**Shows the name of the project assigned to this gene list.
- **Organization**—Shows the name of the organization associated with the gene list. This information cannot be edited; it is derived from the license key.
- **Identifier—**Show the gene identifier.
- **Notes**—Lets you enter information about the genome.

Graph Panel

In the upper right corner of the window is a browser graphing your list. Right-click on the graph for a menu of options. See "Using the Genome Browser" on page 7-1 for information on browser options.

Buttons

The Gene List Inspector window contains the following buttons:

- **Export as Zip—**Saves the gene list as a compressed file.
- **OK**—Saves your data and exit the window.
- **Cancel**—Closes the Experiment Inspector without saving any changes.
- **Help**—Displays online help.
- **Change Projects**—Assigns the selected gene list to a project. Go to "Working with Projects" on page 7-40 for more information.

Gene Lists Tab

The Gene List tab (Figure 12-7) displays a table of all the genes included in the selected list. Double-click a gene or cell in this table to view a Gene Inspector window for the selected gene. See "Using the Gene Inspector" on page 7-16 for information on the Gene Inspector window. Click on any column header in the displayed gene list to sort the table by the values in that column.

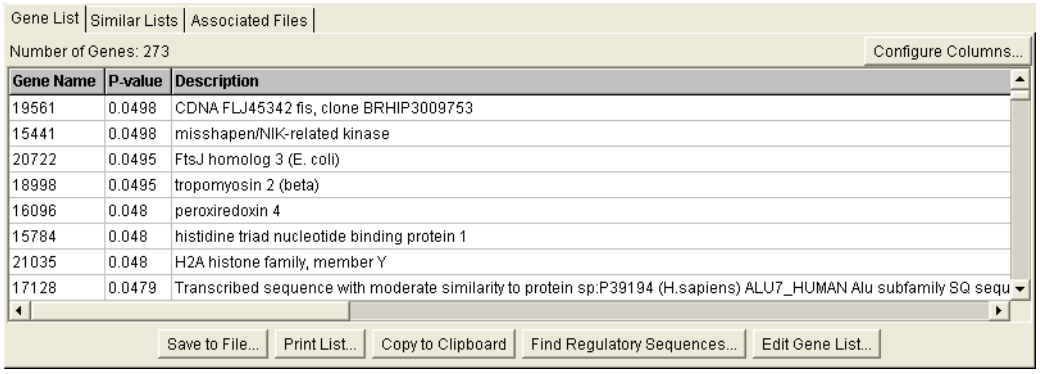

Gene List
Similar Lists
Associated Files
Number of Genes: 273
Configure Columns...
Gene Name
P-value
Description
19561 0.0498 CDNA FLJ45342 fis, clone BRHIP3009753
15441 0.0498 misshapen/NIK-related kinase
20722 0.0495 FtsJ homolog 3 (E. coli)
18998 0.0495 tropomyosin 2 (beta)
16096 0.048 peroxiredoxin 4
15784 0.048 histidine triad nucleotide binding protein 1
21035 0.048 H2A histone family, member Y
17128 0.0479 Transcribed sequence with moderate similarity to protein sp:P39194 (H.sapiens) ALU7_HUMAN Alu subfamily SQ sequ
Save to File...
Print List...
Copy to Clipboard
Find Regulatory Sequences...
Edit Gene List...

Figure 12-7 Gene List Inspector Window: Gene List Tab

From this tab, you have the following options:

- **Configure Columns**—Selects which columns to display in the Gene Lists tab. You can choose from any of the columns in your Master Table of Genes except the Sequence column.
- **Save to File**—Saves the entire gene list as a tab-delimited file.
- **Print List**——Sends the selected gene list to a printer for printing.
- **Copy to Clipboard**—Copies the contents of the gene list to the clipboard. You can then paste the list into another application, such as a text editor or spreadsheet.
- **Find Regulatory Sequences**—Opens the Find Potential Regulatory Sequences window with the current gene list pre-selected. This button is available only if the genome is fully sequenced. For more information on this window, see "Finding Potential Regulatory Sequences" on page 14-1.
- **Edit Gene List**—Opens the Gene List Editor window. For more information, see "Creating or Editing Gene Lists" on page 12-5.

Similar Lists Tab

The Similar Lists tab (Figure 12-8) displays names of lists resembling the selected list, or containing a statistically significant number of overlapping genes. The overlap is calculated using a standard Fisher's exact test and the p-value is adjusted with a Bonferroni multiple testing correction.

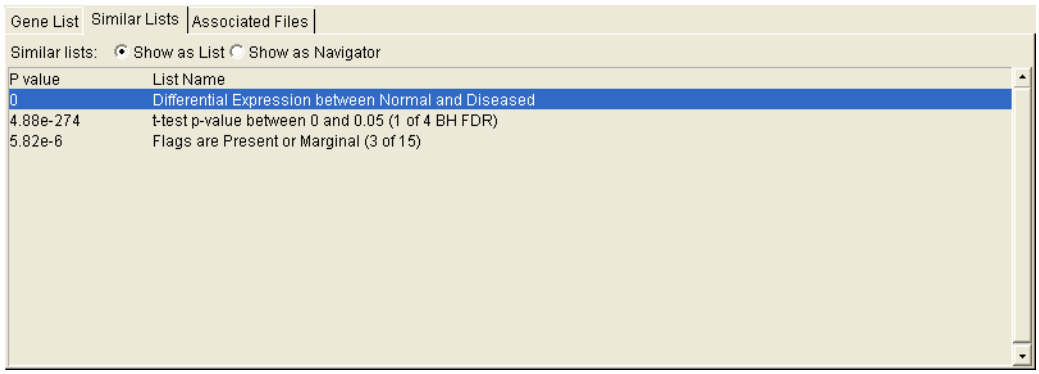


Figure 12-8 Gene List Inspector Window: Similar Lists Tab

Two methods are available to view these lists:

- **List View**—Displays a simple two-column list. In this view, statistical significance is listed as the p-value for each of the similar lists.
- **Navigator View**—Displays a Navigator-style listing.

Right-click a list to print or copy. Double-click a list to view a Gene List Inspector window for that list.

Associated Files Tab

The Associated Files tab (Figure 12-9) lists any files that associated with the selected gene list, such as research publications, documentation, etc.

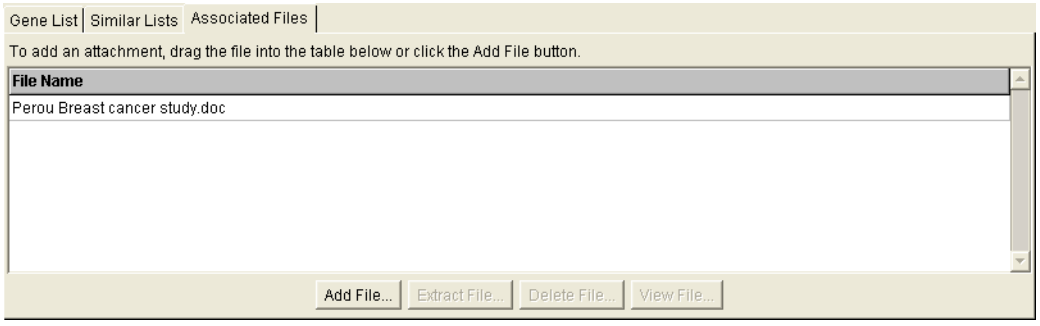


Figure 12-9 Gene List Inspector Window: Associated Files Tab

From this tab, you can perform the following tasks:

- **Add File**—To add a file, click the **Add File** button and select the file you want from the Browse menu that appears. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the file you want from the list and click the **Extract File** button. Choose a location from the Browse menu and click the **Save** button. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click the Delete **button**.
- **View File**—Select a file name in the list and click the **View File** button to view the contents of the file in an external program. The appropriate program is automatically selected if the file type is known.

Opening the Gene List Inspector

To open the Gene List Inspector:

- Do one of the following:
  - In the **Navigator**, open the **Gene Lists** folder, right-click the gene list you want, and then select the **Inspect** command.
  - In the **Navigator**, open the **Gene Lists** folder, and then double-click the gene list you want.

Finding Similar Gene Lists

Similar gene lists in the Gene List Inspector window are gene lists that contain a significant number of overlapping genes with the one selected. The p-value is calculated using the hypergeometric probability. This equation calculates the probability of overlap corresponding to *k* or more genes between a gene list of *n* genes compared against a gene list of *m* genes when randomly sampled from a universe of *u* genes:

$$\frac{1}{\binom{u}{m}} \sum_{i=k}^{n} \binom{m}{i}\binom{u-m}{n-i}$$

The **Use as Standard List** check box in the Gene List Inspector window lets you define a newly created list as standard list. If a list is defined as standard, the list can be included in the automatic tree annotation feature. Go to "Using the Tree View" on page 8-30 for more information about annotating hierarchical trees.

Some lists, such as those created using the Gene Ontology tool, are automatically defined as standard lists.

Configuring Preferences for Gene List Searches

The Restrict Gene List Searches option in the Preferences window (Figure 12-10) lets you limit the lists GeneSpring examines when searching for similar lists in the Gene List Inspector window and during Tree building.

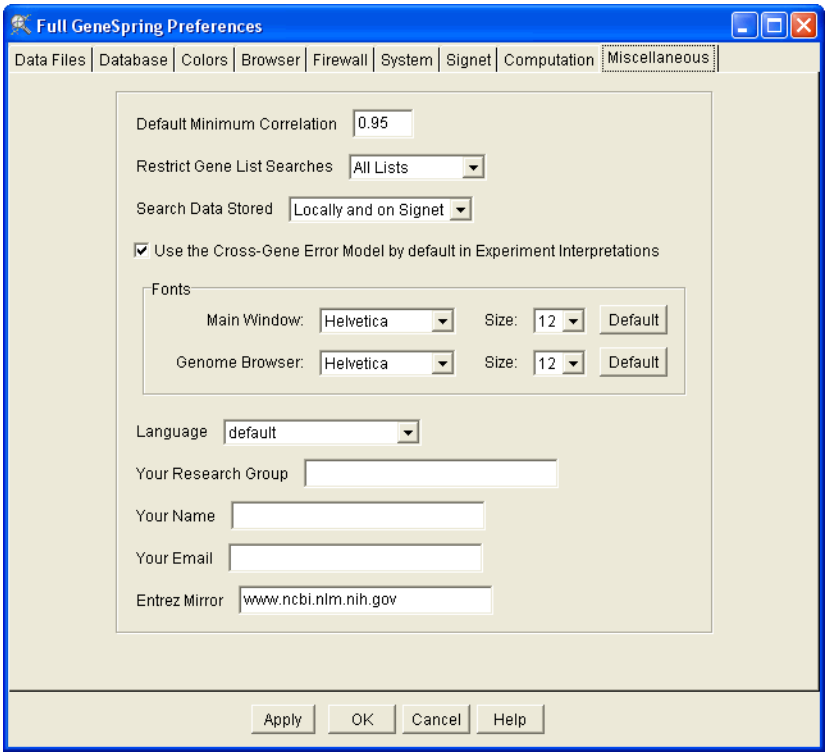


Figure 12-10 Preferences Window: Miscellaneous Tab

To set preferences for gene list searches:

- Select **Edit > Preferences**, select the **Miscellaneous** tab, and change the settings under the **Restrict Gene List Searches** option.

Making Gene Lists

The following section explains how to make gene lists using different functions in GeneSpring. It covers the following topics:

- Making Gene Lists from Annotations
- Making Gene Lists from Venn Diagrams
- Making Gene Lists from Classifications
- Making Gene Lists from Selected Genes
- Making Gene Lists from Expression Profiles

Making Gene Lists from Annotations

Each gene can contain a large number of annotations fields that describe properties of the gene. Annotations can include Description, Product, EC number, GO classifications etc. For certain types of annotations, there are only a limited amount of different values. For example, if an annotation field contains the GO classification of the GO Molecular Process section, there are only a limited amount of GO classes and there are many genes that contain the same value in this annotation field, since they have the same GO classification.

It is possible to create gene lists in GeneSpring that group all genes together with the same value in one of the annotation fields. The gene list will be created by comparing the values in the annotation column, and creating as many gene lists as there are unique values. Each gene that contains one of these values, will be placed in the gene list so that each gene in those gene list will share the same annotation.

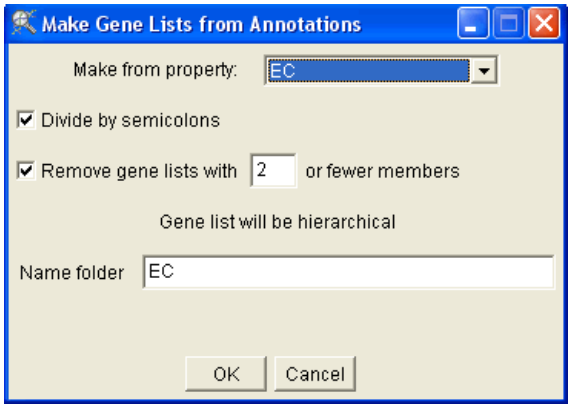
This is particularly useful when the annotation fields contains some sort of classification, like GO classification or other classifications.

For more information about annotations, go to Chapter 11, “Annotating Genes”.

To make gene lists from annotations:

1. In the **Navigator**, open the **Gene Lists** folder, and then select the gene list you want.
2. Select **Annotations > Make Gene Lists from Annotations**.

   The Make Gene Lists from Annotation window opens.



3. In the **Make from Annotations** menu, choose the annotation you want to use for generating gene lists.

   Go to "Master Table of Genes Editor Window" on page 6-36 for a description of the annotations.

   After making a selection, the **Name Folder** field is automatically filled in for you. Alternatively, you can change the folder name if you want.

   Note: If the annotation column you chose does not have any values, a warning message will appear stating "No annotations available for this property". If you continue and press OK, you will get another warning message and no gene lists will be created.

4. Do one of the following:

   - Some annotations fields contain multiple distinct values that are separated by semicolons. If you want to create a gene list for each of the individual values, select the **Divide by Semicolons** check box.
   - If you want the complete cell value to be regarded as one value, leave the **Divide by Semicolons** check box unchecked.

5. To exclude the creation of gene lists that would only have a few members, select the **Remove gene lists with 2 or fewer members** check box.

   It is usually not meaningful to have gene lists with only a few members, since this means that there are only a few genes with these same annotations. Use this option to exclude those gene lists. This feature is enabled by default and will exclude gene lists with 2 or fewer members.

6. Click the **OK** button.

After completion, a new folder will be created in the Gene Lists folder. By default, it will have the same name as that used by the annotation column. The folder will contain the gene lists, along with names that are the values from the annotation column. Each gene lists will contain genes that share the same value in the annotation column.

Note: It is possible to create thousands of different gene lists with this feature and care should be taken in choosing the annotation column. If thousands of gene lists are created, some of the features of GeneSpring can take some time to complete, such as the Find Similar Lists function.

Making Gene Lists from Venn Diagrams

A Venn diagram lets you quickly visualize genes common to more than one gene list. You can also find genes present only in a particular list. The gray area behind the circles represents the Venn diagram "universe" (the selected gene list). Genes in the selected list that are common to gene lists represented by the Venn diagram circles appear as numbers in those circles.

To make a gene list from a Venn Diagram:

1. Create a Venn diagram of the gene lists you want to compare.

   For information about creating and filling Venn diagrams, see "Coloring by Venn Diagram" on page 8-64.

2. Right-click the area of the Venn diagram in which you want to make a list.

3. Select **Make List of Genes in This List Only**.

   The New Gene List window opens.

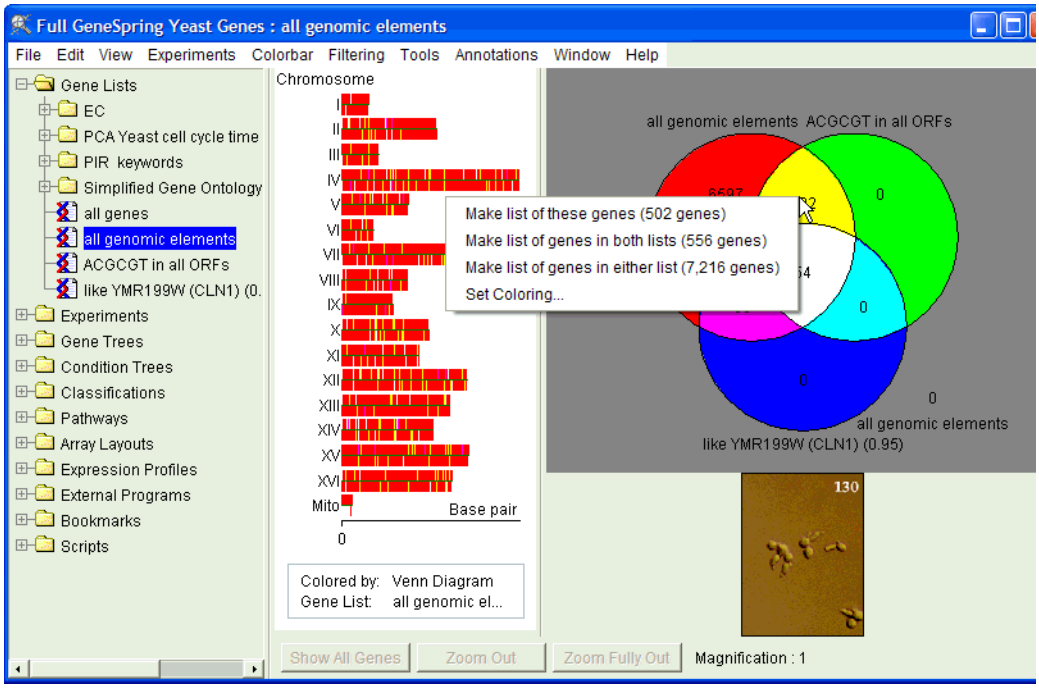
   When you click in an area where two circles overlap, you have the following options:

   - **Make list of these genes—**Lists genes in the immediate geometric area.
   - **Make list of genes in both lists—**Lists genes common to the two circles; for example, the intersection.
   - **Make list of genes in either list—**Lists all genes in the two circles; for example, the union.

   If you click in an area where three circles overlap, you have the following options—

   - **Make list of genes in all lists—**Lists genes common to the three circles; for example, the intersection.
   - **Make list of genes in any list—**Lists all genes in the three circles; for example, the union.

4. If you click a non-overlapping (gray) area in a circle, you can make a list of genes in this list only.



5. Specify the gene list from which to obtain associated numbers and click the **OK** button.

   The resulting gene list can “inherit” the associated values from the gene list that is used to create the Venn diagram. If there is only one gene list with associated numbers, those numbers are used. If two or more of the gene lists contain associated values, the user is offered the choice which associated values to use. A dialog box will appear with the names of the two or three gene lists that contain associated values and you can choose which one to use with a radio button. you want to use.

6. Name and save your new list.

Making Gene Lists from Classifications

A classification is a division of a selected list of genes into several groups. It is usually created when performing a k-means, SOMs or QT clustering, and is usually saved under the Classification folder. No gene can be in more than one group.

You can generate gene lists from any classification. For example, if you have a five cluster k-means classification, you can view which genes are in each cluster by making a gene list from the k-means classification.

A classification can be saved as a gene list, but gene lists cannot be saved as classifications.

Classifying Genes into Groups

You can use gene lists to classify genes into several groups. In this case, a single gene can be found in more than one group. Groups of gene lists can also be used to color genes, similar to a classification. For that purpose, gene lists must be placed into subfolders under the Gene List folder. Then, right-click on the gene list folder and select **Use as Coloring** or **Split Window > Both**.

Creating Your Own Classifications

You can create your own classification and paste it in GeneSpring. It will be saved under the Classification folder in the Navigator.

Note: Only one gene is allowed per class. If you wish to import overlapping groups of genes, import each gene list separately and organize them into a folder

To create a classification:

1. Open an Excel spreadsheet.
   a. Format the list of genes you wish to import as classification as follows:
   b. In the first cell, enter the word “Classification”.
   c. In the cell on the right of “Classification”, enter the name of the classification as it will appear in GeneSpring.
   d. In the column below “Classification”, enter the list of Gene Identifiers that will be part of the classification. Use the same type of Gene Identifiers as used in your genome. To find out what type of identifier is used in your genome, double-click on the “All Genes” list and look up the “Systematic” names.
   e. In the column below the classification name, enter the name of the class or group, as in the example below:

Classification	Functional classes
L1	Cell adhesion
GAP43	Cell growth
NGF	Cell growth
trkC	Cell growth
PTN	Cell adhesion
GDNF	Cell growth
synaptophysin	Transport
GAT1	Transport
GRa3	Transport

2. Copy the classification from the spreadsheet.
3. Go to GeneSpring and select **Paste Classification** from the **Edit** menu.
4. To save the classification as a gene list, do one of the following:
   - In the Choose Classification Name window, select **Gene Lists**.

- Right-click the classification in the Classification folder and select **Make Gene Lists** from the pop-up menu.

Making Gene Lists from Classifications

To make a gene list from a classification:

1. In the **Navigator**, open the **Classifications** folder.
2. Right-click the classification you want and select **Make Gene Lists**.

   GeneSpring creates a gene list folder for the classification containing one list for each cluster. This folder appears in the Gene Lists folder in the Navigator.

Making Gene Lists from Selected Genes

The Make List from Selected Genes command lets you make lists from genes you select graphically.

To make a gene list from selected genes:

1. In the **Genome Browser**, do one of the following:
   - Hold down **Shift**, click a region, and drag a rectangle across the area you want to select. Release the mouse button before releasing **Shift**.
   - Select multiple genes by clicking over their representative lines or rectangles while holding down Shift.

     Selected genes appear in the color you specified in the Preferences window.

     Go to "Colors Tab" on page 3-17 for more information.
2. Once you have selected all the genes you want, right-click in the Genome Browser and select **Make List from Selected Genes**.

   The New Gene List window opens.
3. Name your list and click the **Save** button.

   For more information about this window, see "Gene List" on page 12-1.

Making Gene Lists from Expression Profiles

The Creating Expression Profiles function lets you draw a pseudo-gene to represent a hypothetical expression pattern. This function is useful if you have some idea of what gene expression pattern you are looking for, as you can simply draw a pattern and look for genes that behave similarly.

You must be in Graph view to create an expression profile. Double-click the expression profile to open the Gene Inspector for that gene.

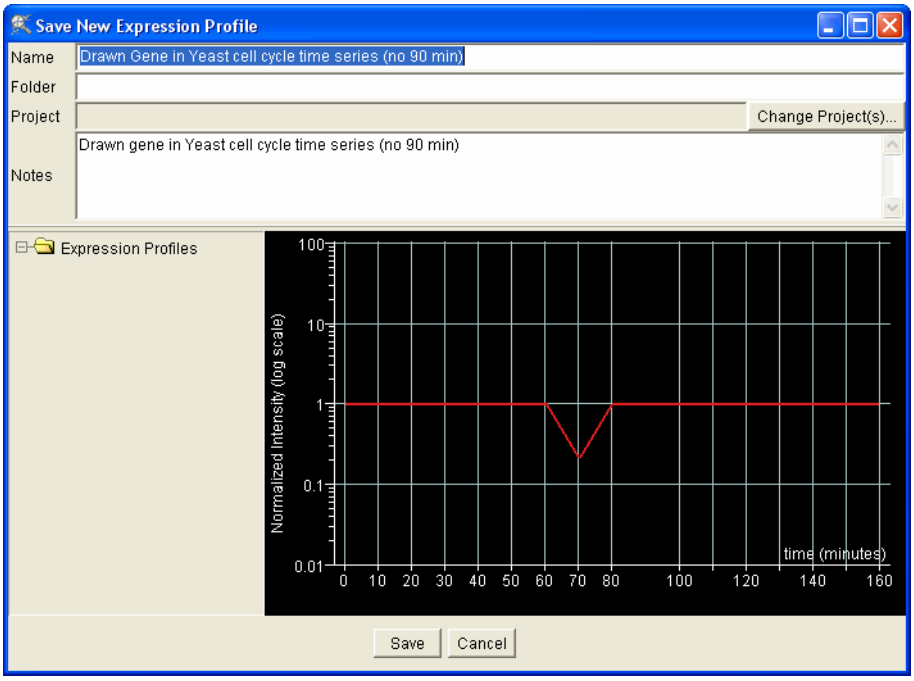
To create an expression profile:

1. In the **Navigator**, open the **Gene Lists** folder.
2. Select the gene list you want.
3. Select **View > Graph**.

4. Select the gene you want.
5. Select **Tools > Draw Expression Profile**.

   A new gene appears on the window at the normalized median of your data (usually 1.0).
6. Hold down the **Ctrl** key (**Alt** on Macintosh).
7. Click and drag the mouse in the desired shape.
8. Right-click the shape and select the **Save Expression Profile** command.

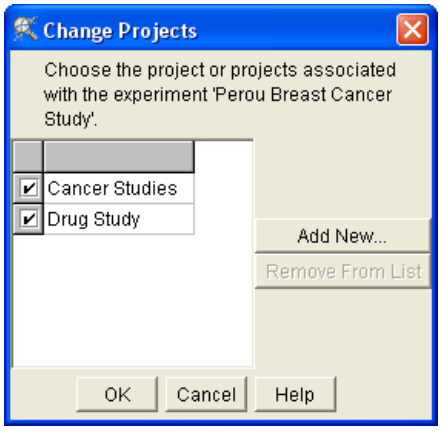
   The Save Expression Profile window opens.



9. In the **Name** field, enter a name for the expression profile.

   Be sure to choose a descriptive name that you will remember later.
10. To save the expression profile in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.
11. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.
12. To assign the expression profile to a project, click the **Change Projects** button.

The Change Projects window opens.



13. Do one of the following:

- To create a new project name, do the following:
  a. Click the **Add New** button.

     The Add New window opens.

  b. In the **Project Name** field, enter a name for the project.

     All data objects in the project will be associated with this project name.

  c. Click the **OK** button.
- To assign a previously created project name, do the following:
  a. Select the project name(s) you want.

     You can assign data objects to more than one project.

  b. Click the **OK** button.

14. Click the **Save** button.

Your new expression profile appears in the Expression Profiles folder in the Navigator.

Working with Homology Tables

Homology tables allow you to compare the expression results from one genome or array to the results from a different genome or array. Once an homology table is created, you can "translate" the gene list or experiment from one genome to a gene list of the corresponding genes in the other genome. This is useful if you have performed similar experiments on different chips or technologies and want to compare the behavior of a set of genes on the two technologies.

The Homology tool automates the process of building homology tables for certain organisms. Currently, a limited list of organisms is available. Using this tool, homologies can be made between any pair of organisms that are included in both HomoloGene and UniGene. Within-genome homologies are based solely on UniGene Cluster ID or LocusLink Locus ID.

This section covers the following topics:

- Build Homology Tables Window
- Supported Organisms
- Homology Table Examples
- Building Homology Tables
- Viewing Homology Tables
- Deleting Homology Tables

Build Homology Tables Window

The Build Homology Tables window (Figure 12-11) automates the process of building homology tables for certain organisms. Currently, a limited list of organisms is available. Using this tool, homologies can be made between any pair of organisms that are included in both HomoloGene and UniGene. Within-genome homologies are based solely on UniGene Cluster ID or LocusLink Locus ID.

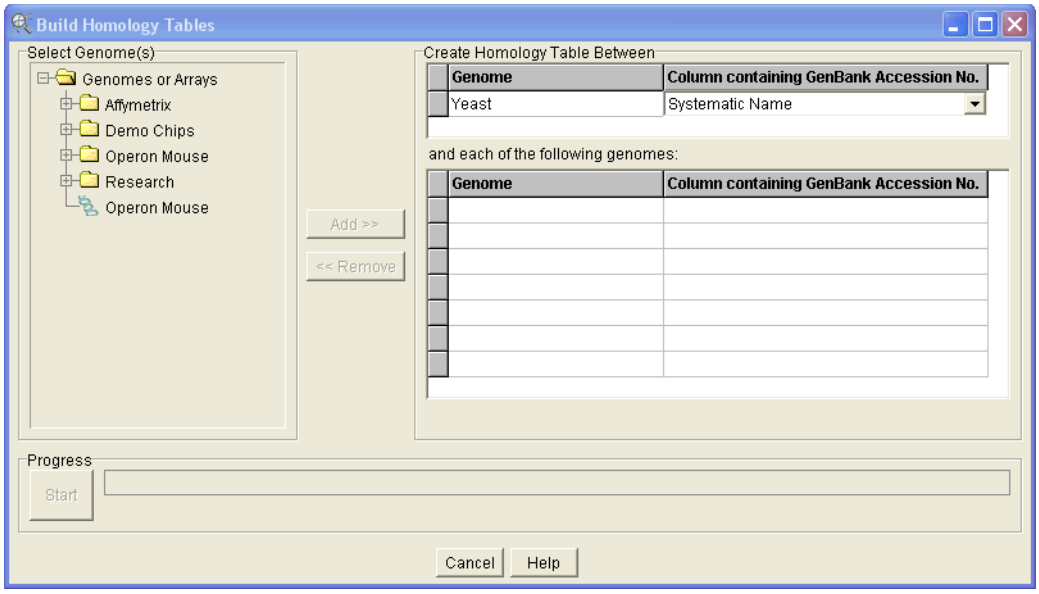


Figure 12-11 Build Homology Tables Window

This window contains the following elements:

Select Genomes Navigator

The Select Genomes box displays the genomes and arrays currently available on GeneSpring and Signet. Double clicking a genome adds it to the lower table in the window.

Buttons

- **Add**—Adds the selected genome to the lower table.
- **Remove**—Removes the selected genome from the lower table.
- **Start**—Starts the process of building the homology tables.

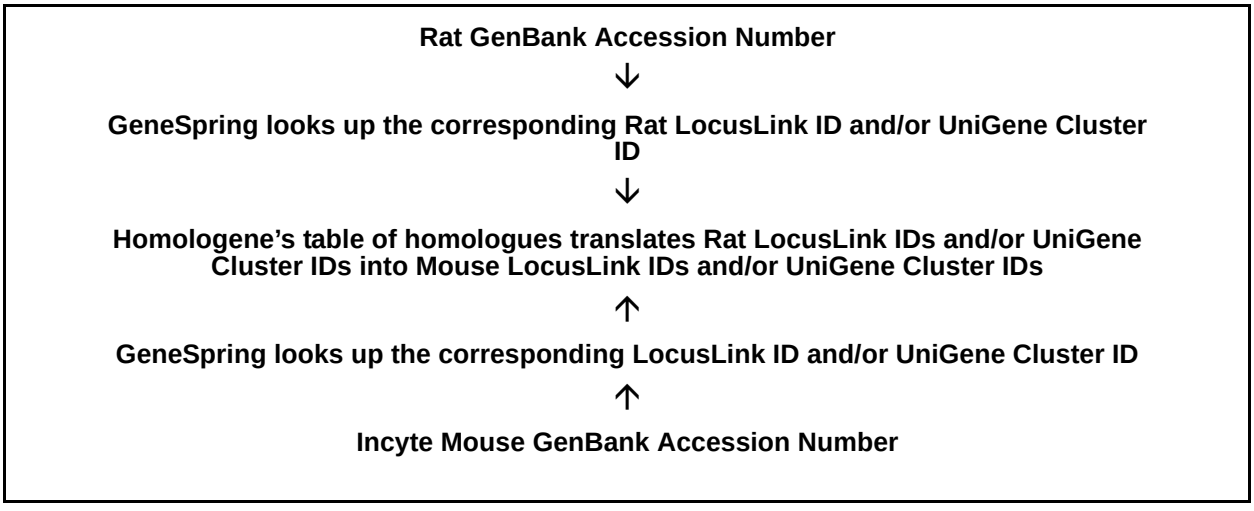
Supported Organisms

Homologies can be created for the following organisms:

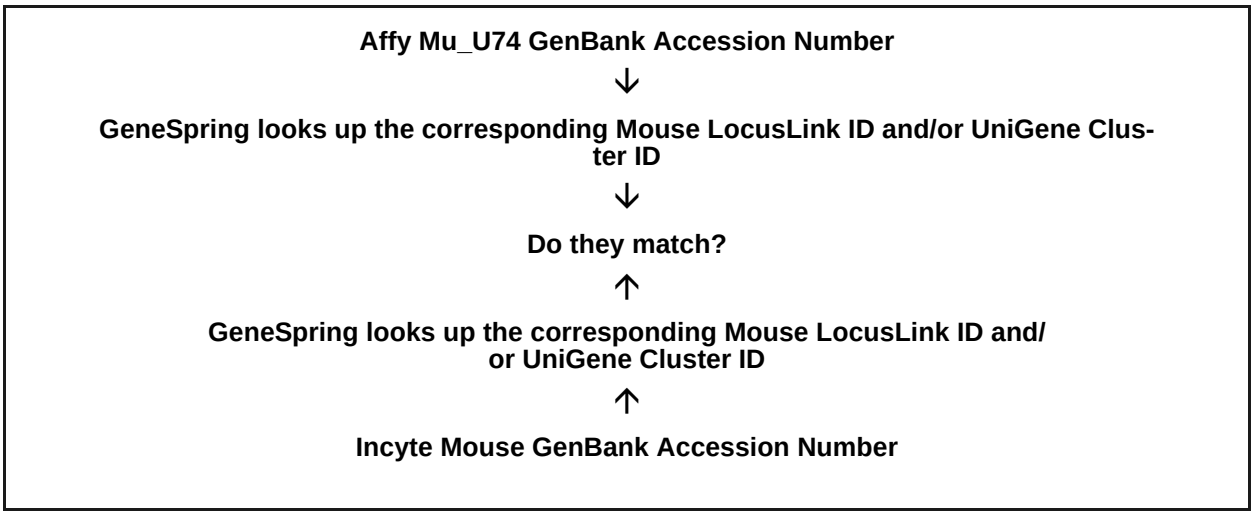
- Cow (Bos Taurus)
- Nematode (Caenorhabditis elegans)
- Sea squirt (Ciona intestinalis)
- Zebrafish (Danio rerio)
- Fruit fly (Drosophila melanogaster)
- Human (Homo sapiens)
- Mouse (Mus Musculus)
- Rat (Rattus norvegicus)
- Pig (Sus scrofa)
- Clawed frog (Xenopus laevis)
- Thale cress (Arabidopsis thaliana)
- Barley (Hordeum vulgare)
- Rice (Oryza sativa)
- Wheat (Triticum aestivum)
- Maize (Zea mays)

Homology Table Examples

The following example show how GeneSpring builds a homology table between Rat GenBank Accession Numbers and Mouse GenBank Accession numbers.


Rat GenBank Accession Number
↓
GeneSpring looks up the corresponding Rat LocusLink ID and/or UniGene Cluster ID
↓
Homologene's table of homologues translates Rat LocusLink IDs and/or UniGene Cluster IDs into Mouse LocusLink IDs and/or UniGene Cluster IDs
↑
GeneSpring looks up the corresponding LocusLink ID and/or UniGene Cluster ID
↑
Incyte Mouse GenBank Accession Number

The following example shows how GeneSpring makes a homology table between Affy Mu_U74 GenBank Accession Numbers and Incyte Mouse GenBank Accession Numbers.


Affy Mu_U74 GenBank Accession Number
GeneSpring looks up the corresponding Mouse LocusLink ID and/or UniGene Cluster ID
Do they match?
GeneSpring looks up the corresponding Mouse LocusLink ID and/or UniGene Cluster ID
Incyte Mouse GenBank Accession Number

Building Homology Tables

To build a homology table:

1. Do one of the following:
   - Select **Annotations > Build Homology Tables**.
   - Open the **Genome Inspector**, click the **Homology Tables** tab, and click the **Add/Update** button.

     Go to "Opening the Genome Inspector" on page 6-20 for instructions on opening the Genome Inspector.

     The Build Homology Tables window opens.
2. From the menu in the **Column Containing GenBank Accession No.**, select the appropriate column that contains the GenBank (or EMBL) accession number for the selected genome.
3. In the **Navigator**, select the genome you want and click **Add**.

   The genome is added to the lower table on the right side of the window.
4. From the menu next to the newly added genome, select the name of the column containing the genome's GenBank Accession Number.
5. Repeat steps 2 through 4 for each genome you want to add.
6. Click the **Start** button.

   This process can take several hours to complete. If the initially selected genome does not have GenBank Accession Numbers, an error message appears. If a selected genome is not on Homologene, you will receive an error message after the Homology tool has finished running.
7. When prompted, specify whether or not to save the UniGene Cluster IDs.

When you choose to select the UniGene ID's, they will be saved in a annotation column called "Unigene". Any data that was previously in the Unigene column will be overwritten.

Viewing Homology Tables

To view a homology table:

1. Open the **Genome Inspector** and click the **Homology Tables** tab.

   Go to "Opening the Genome Inspector" on page 6-20 for instructions on opening the Genome Inspector.

2. Select the homology table you want to view.
3. Click the **View Homologs** button.

   The Homology Table Viewer window opens.

4. To display only those genes with homologs, select the **Only show genes with homologs** option.
5. To copy the entire homology table to the clipboard, do the following:
   a. Click the **Copy to Clipboard** button.
   b. Paste the data in the target file.
6. To add columns to the homology table, do the following:
   a. Click the **Configure Column** button.
7. To configure the columns, do the following:
   a. Click the **Configure Columns** button.

      The Configure Columns window opens.

      Within each section of the Configure Columns window the annotations are listed alphabetically. The order shown here is the same order in which the annotations are displayed in the Edit Genes and Annotations window. This window contains the following elements:

      - **Identifier Columns**—Shows the annotations GeneSpring always has in memory. They will be available for every genome, even if they are empty.
      - **Standard Columns**—Shows the annotations that the GeneSpider uses (except for User Notes). These may or may not exist for a given genome.
      - **Custom Columns**—Shows all the columns the user has defined themselves.
      - **Other Information**—Contains those "annotations" that aren't really annotations, but are useful to have.

   b. To select all options, click the **Check All** button.
   c. To clear all options, click the **Clear All** button.
   d. To select individual options, select the options you want.
   e. Click the **OK** button.

8. Click the **Close** button.

Deleting Homology Tables

To delete a homology table:

1. Open the **Genome Inspector** and click the **Homology Tables** tab.

   Go to “Opening the Genome Inspector” on page 6-20 for instructions on opening the Genome Inspector.

2. Select the homology table you want to delete.
3. Click the **Delete** button.
4. Click the **Close** button.

13

Finding Differentially Expressed Genes

This chapter explains how to find genes that are differentially expressed between conditions using analysis of variance (ANOVA). It covers the following topics:

- Statistical Analysis (ANOVA)
- Before You Begin
- Performing One-way ANOVA
- Performing Two-way ANOVA
- Interpreting ANOVA Results
- Viewing Generated P-values

Overview

GeneSpring's robust statistical tools give researchers flexibility in designing complex experimental structures, and provide confidence in analyzing results. With analysis of variance (ANOVA) techniques, you can identify differentially expressed genes when the groups being compared are defined by one or two parameters. GeneSpring provides in-depth techniques for analyzing and visualizing the variability of genes differentially expressed across independent groups. These techniques include:

- One-way -and two-way ANOVA
- Multiple Testing Corrections
- False Discovery Rate Prediction
- Tukey and Student-Newman-Keuls post hoc tests

Statistical Analysis (ANOVA)

The purpose of analysis of variance is to test for a significant difference in expressed levels between conditions. In GeneSpring, Statistical Analysis (ANOVA) is a filter tool that statistically compares mean expression levels between two or more groups of samples. The object is to find the set of genes for which the specified comparison shows statistically significant differences in the mean normalized expression levels as interpreted according to your current interpretation mode (logarithm, ratio, or fold change) across all the groups. This comparison is performed for each gene, and the genes with the most significant differential expression (smallest p-value) are returned.

Filtering genes based on a one-sample t-test of the mean expression level across repeats or replicates versus a reference value can be done by selecting “t-test p-value” as the filter criteria in Expression Percentage Restriction.

One-Way ANOVA

One-way analysis of variance lets you determine if one given factor, such as drug treatment, has a significant effect on gene expression behavior across any of the groups under study. A significant p-value resulting from a one-way ANOVA test would indicate that a gene is differentially expressed in at least one of the groups analyzed.

If there are more than two groups being analyzed, the one-way ANOVA does not specifically indicate which pair of groups exhibit statistical differences. Post hoc tests can be applied in this specific situation to determine which specific pairs are differentially expressed.

The one-way ANOVA test filters out genes that do not vary significantly across different groups with multiple samples. This analysis lets you find those genes that exhibit important changes between various conditions of the experiment. This comparison is performed for each gene, and the genes with sufficiently small p-values are returned.

Comparisons can be performed with parametric or non-parametric methods. The parametric comparison for two groups is known as Student’s two-sample t-test. For multiple groups, this is known as one-way ANOVA. You can specify whether to assume within-group variances are equal variances across all groups.

Calculations without the assumption of equality of variances are done using Welch’s approximate t-test and ANOVA. Non-parametric comparisons are also available, corresponding to the Wilcoxon two-sample text (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups.

In GeneSpring, this type of analysis is called the “1-way ANOVA.” Go to “Performing One-way ANOVA” on page 13-5 for more information.

Two-Way ANOVA

The two-way analysis of variance is an extension to the one-way analysis of variance. In two-way ANOVA, two independent parameters are used in the analysis. The two-way ANOVA tests genes for significant differences across groups defined by two parameters. This test is appropriate to use for a two-way design where the groups to be compared are defined by two parameters.

An example of a two-way design is an experiment in which you have two groups of samples; such as cancer cells and normal cells. You also split each of the two groups in two groups again, and treat one of these group with some drugs, while treating the other group only with some control. Given this design, there are two independent parameters, “Disease state” (with values “Cancer” and “Normal”) and “Treatment” (with values “Treated” and “Untreated”).

Using two-way ANOVA, you can determine if there are any genes that are differentially expressed between cancer and normal cells. Moreover, you can determine if there are any genes that are differentially expressed due to the treatment condition. You can also determine if there is any interaction between the two parameters; that is, if any of the genes show a different behavior towards the treatment, depending on whether they are treated or not.

A two-way design is one that can be thought of as a matrix, with the rows indexed by the values of one parameter, and the columns indexed by the values of a second parameter. Each cell then represents the number of replicates in that particular group. Ideally, each cell should have an equal number of replicates. This is called a *balanced design*. GeneSpring can also perform two-way ANOVA for proportional designs. You will not be able to perform two-way ANOVA for unbalanced designs that are not proportional.

Performing a two-way ANOVA tests for the effect of each parameter, as well as the interaction between them, simultaneously. For each gene, three p-values are produced, one for each parameter and one for the interaction term. Genes with p-values less than the specified cutoff are returned. From the results window, several options for creating gene lists are available.

In GeneSpring, this type of analysis is called the “2-way ANOVA.” Go to “Performing Two-way ANOVA” on page 13-15 for more information.

Statistical Analysis (ANOVA) Window

The Statistical Analysis (ANOVA) window (Figure 13-1) lets you perform one-way ANOVA, two-way ANOVA, post hoc testing, and multiple testing correction.

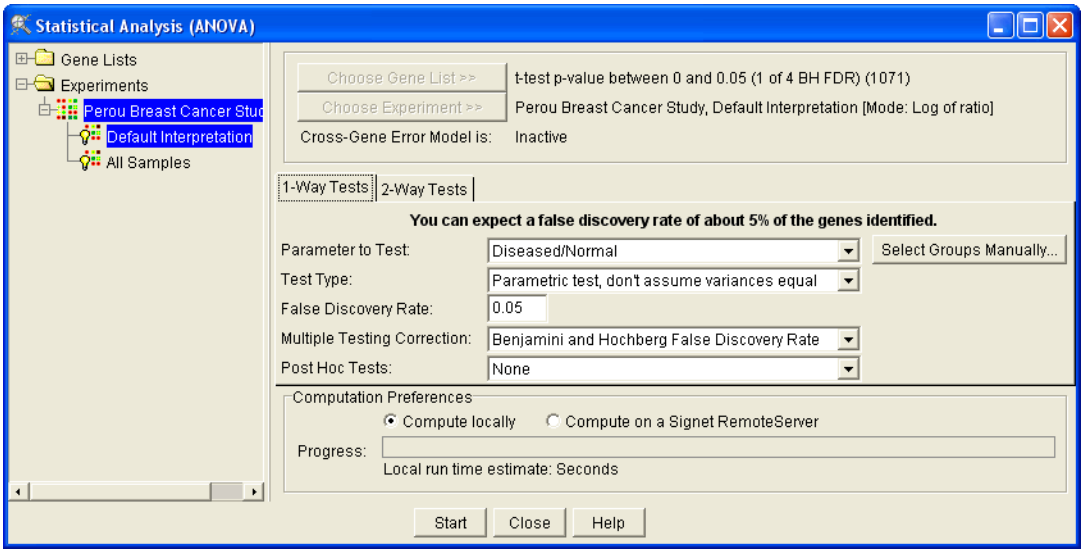


Figure 13-1 Statistical Analysis (ANOVA) Window

This window contains the following elements:

Choose Gene List Button

The Choose Gene List button lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

Choose Experiment Button

The Choose Experiment button lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be the log-of-ratio mode.

Parameter to Test Option

The Parameter to Test option lets you select the parameter and the underlying groups to compare. If you want to compare only selected conditions for this parameter, open the **Select Groups Manually** window, and clear the conditions that you would like to ignore. Only groups that are checked will be analyzed. Ensure that your data has been log-transformed (by selecting log-of-ratio mode in Experiment Interpretation window).

P-value cutoff or False Discovery Rate Option

The False Discovery Rate option indicates the overall rate of false positives. Also known as a Type I Error, common false discovery rates are 0.05 or 0.01. The wording for this option, and its final effect on the number of false positives, changes according to the multiple testing correction selected.

Multiple Testing Corrections Options

Since you will be performing a large number of individual tests, the ANOVA can suffer from the so-called “multiple testing” problem. even if you only select those genes with a p value of less than 0.05, in a typical experiment with 10,000 genes, you could still select 500 genes at random that you will consider to be significantly differentially expressed in your experiment.

Multiple testing correction will allow you to correct for this problem. GeneSpring offers a number of different multiple testing correction options, and each option has its advantages and disadvantages. See the section “Multiple Testing Corrections” on page 13-12 for more informations.

Post Hoc Tests

Post hoc tests let you determine which pair(s) among the groups under study have expression means that are statistically significant. For example, performing a one- or two-way ANOVA can give you a list of genes that are differentially expressed. If you have more than two groups, however, you will not be able to determine in which groups the genes are differentially expressed. Post hoc tests will allow you to determine which of the groups actually show the differentially expressed genes.

Before You Begin

Before you perform an ANOVA, ensure that you have met the following guidelines:

1. Do you have replicates for the experimental groups that you are about to compare? Statistical tests that compare one group to another, such as Student's t-test or one-way ANOVA, need variance and means for each group.

   Without replicates, the variance for each group cannot be computed using standard methods. However, variance for experimental groups without replicates can be computed by applying the GeneSpring Cross-Gene Error Model. If no replicates are available, apply the Error Model based on Deviation from 1 before proceeding.

   Refer to the online technical notes, Webinars, or Cross-Gene Error Model features sheet on http://www.silicongenetics.com for more information.
2. Have you filtered out genes whose measurements are mostly unreliable?
3. Have you defined one parameter in the Experiment Parameters window indicating which sample belongs to which group?
4. If you plan to use a parametric test, have you changed the analysis mode to log of ratio in the Experiment Interpretation window? Parametric tests assume that means of the populations under study are normally distributed (Gaussian distribution). Interpreting your data in log mode will make data more Normal/Gaussian than ratio mode.
5. It is mandatory that you either have replicates or apply the Cross-Gene Error Model if no replicates are available, in order to perform one-way ANOVA for groups under study. It is also recommended (though not mandatory) that your statistical analysis be performed on a set of reliable genes, instead of all genes, on the chips.

Performing One-way ANOVA

This section explains how to perform a one-way ANOVA. The following topics are covered:

- Performing One-way ANOVA
- Technical Details for One-way ANOVA
- Multiple Testing Corrections
- Post Hoc Tests

Performing One-way ANOVA

To perform one-way ANOVA:

1. Select **Tools > Statistical Analysis (ANOVA)**.

   The Statistical Analysis (ANOVA) window opens.
2. Click the **1-way Tests** tab.

3. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and click the **Choose Gene List button**.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

4. In the **Navigator**, select the experiment interpretation you want to analyze, and click the **Choose Experiment** button.

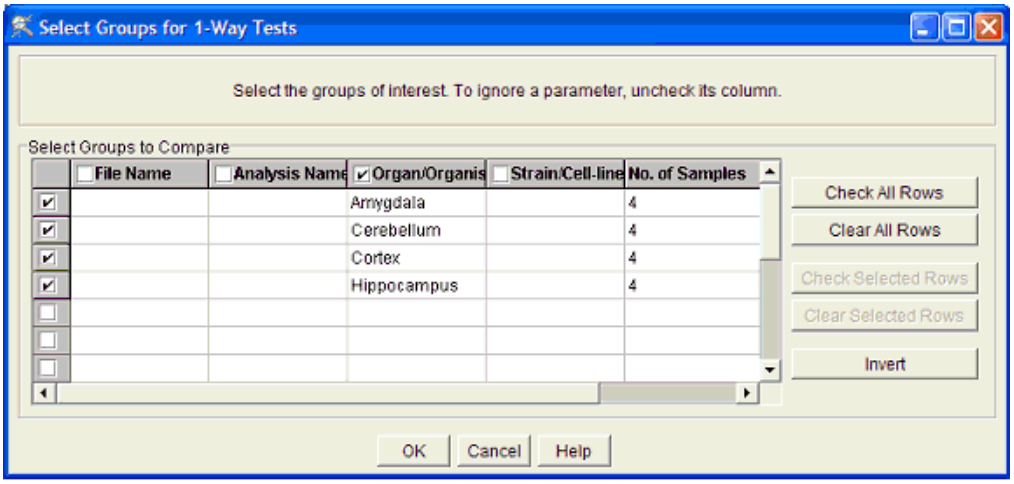
   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

5. In the **Parameter to Test** list, do one of the following:

   - Select the parameter that defines the group.
   - To compare only selected conditions for this parameter, go to step 6.

6. Click **Select Groups Manually**.

   The Select Groups for 1-Way Tests window opens.



   a. Clear the box in a column's header to ignore groupings based on that parameter.

      Only groups that are checked will be analyzed.

      When you perform this step, the table is dynamically updated to reflect the change. The number of rows decreases and the number of samples associated with each condition increases.

   b. Define the conditions to compare by selecting or clearing the check box in a given row.

      The **Check All/Clear All** buttons allow you to select or clear all rows. The **Invert** button selects all unchecked rows, and clears all selected rows.

7. In the **Test Type** list, select the type of test to perform. There are four testing options:
   - **Parametric test, assume variances equal**—Filter based on the results of a Student's two-sample t-test for two groups or a one-way analysis of variance (ANOVA) for multiple groups.
   - **Parametric test, don't assume variances equal**—Filter based on the results of an ANOVA or Welch's approximate t-test for two groups. This is the most appropriate test for standard experiments in which the global error model is not turned on or should not be used in the analysis.
   - **Parametric test, use all available error estimates**—Filter based on the variances estimated by the Cross-Gene Error Model. If the Cross-Gene Error Model is not turned on, this test is equivalent to the Parametric test.
   - **Non-Parametric test**—Filter based on the rank of each sample, rather than the expression level. Non-parametric comparisons use the Wilcoxon two-sample rank test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups. This test is most successful if you have more than five replicate samples in each group.
8. Select a **P-value cutoff** for genes that pass the filter.

   The p-value indicates a probability, with a value ranging from zero to one, that the difference observed between groups is due to chance. The lower the p-value, the more significant the difference between the groups.

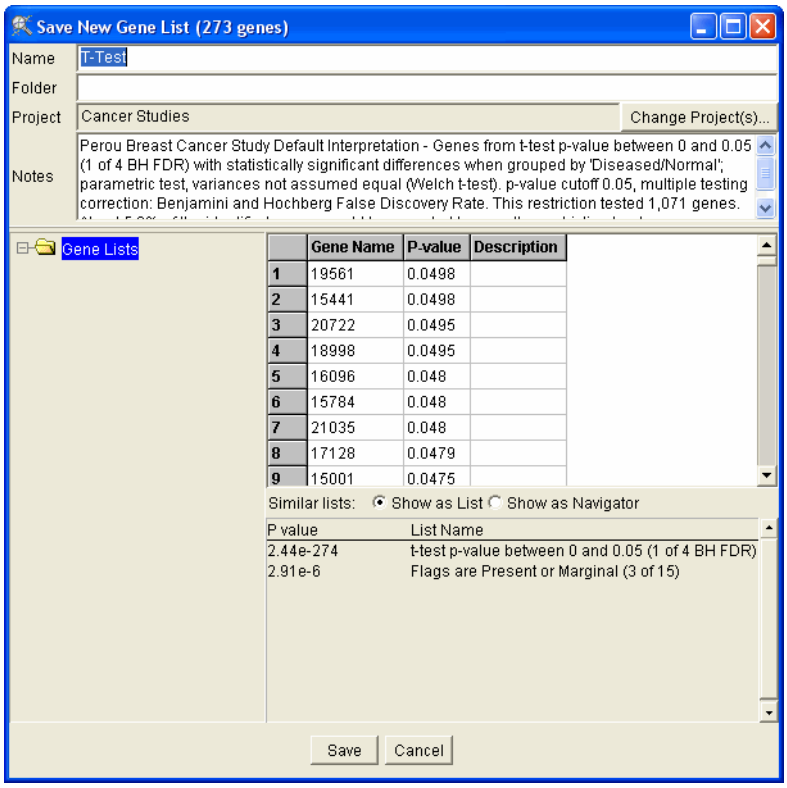
   This field will actually change to **False Discovery Rate** when you choose one of the Multiple Testing Correction options.
9. Select a type of multiple testing correction.

   Go to "Multiple Testing Corrections" on page 13-12 for a description of the options.
10. (Optional) Select a type of post hoc test to perform.

    Go to "Post Hoc Tests" on page 13-13 for a description of the tests.
11. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

    This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.
12. Click the **Start** button.

    As the ANOVA executes, you can watch its progress in the Progress bar. Results are shown in the Save New Gene List window.


Save New Gene List (273 genes)
Name
T-Test
Folder
Project
Cancer Studies
Change Project(s)...
Notes
Perou Breast Cancer Study Default Interpretation - Genes from t-test p-value between 0 and 0.05 (1 of 4 BH FDR) with statistically significant differences when grouped by 'Diseased/Normal'; parametric test, variances not assumed equal (Welch t-test). p-value cutoff 0.05, multiple testing correction: Benjamini and Hochberg False Discovery Rate. This restriction tested 1,071 genes.
Gene Lists
Gene Name
P-value
Description
1 19561 0.0498
2 15441 0.0498
3 20722 0.0495
4 18998 0.0495
5 16096 0.048
6 15784 0.048
7 21035 0.048
8 17128 0.0479
9 15001 0.0475
Similar lists:
Show as List
Show as Navigator
P value
List Name
2.44e-274
t-test p-value between 0 and 0.05 (1 of 4 BH FDR)
2.91e-6
Flags are Present or Marginal (3 of 15)
Save
Cancel

13. In the **Name** field, enter a name for the gene list.

    Be sure to choose a descriptive name that you will remember later.

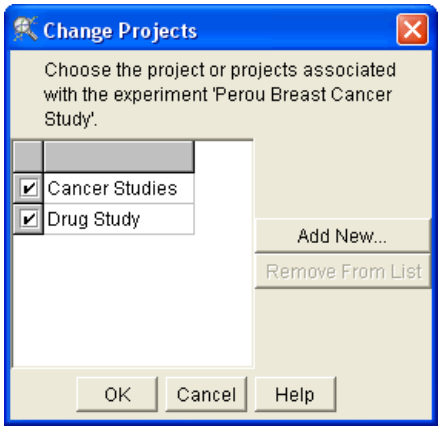
14. To save the gene list in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window, and leave the **Folder** field blank.

15. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.

16. Enter **Notes** containing more descriptive information about the gene list.

17. To assign the gene list to a project, click the **Change Projects** button.

    The Change Projects window opens.



18. Do one of the following:
    - To create a new project name, do the following:
        a. Click the **Add New** button.

           The Add New window opens.

        b. In the **Project Name** field, enter a name for the project.

           All data objects in the project will be associated with this project name.

        c. Click the **OK** button.
    - To assign a previously created project name, do the following:
        a. Select the project name(s) you want.

           You can assign data objects to more than one project.

        b. Click the **OK** button.
19. Click the **Save** button.
20. Go to "Interpreting ANOVA Results" on page 13-23.

Technical Details for One-way ANOVA

This section describes the different computations that GeneSpring performs in one-way ANOVA.

Computations for Genes

GeneSpring performs the following computations for each gene in the analysis:

Let i index over the G groups formed by distinct levels of the comparison parameter.

Let X_{ik} be the expression values, with k running over the replicates for each situation, interpreted according to the current interpretation (ratio, log of ratio, fold change).

Let:

N_i = the number of non-missing data values for each group,

$\bar{X}_i = 1/N_i \sum_{k-1}^{N_i} X_{ik}$ be the group means, and

$SS_i = \sum_{k-1}^{N_i} (X_{ik} - \bar{X}_i)^2$ be the within-group sum of squares.

In all calculations, missing values (No Data) or (NaN) are left out of the sums, not propagated. If any of the N_i are zero, drop that parameter level from the analysis, and readjust G accordingly. If G is not at least 2, exit (p-value = 1).

Computations for Parametric Test, Variances Assumed Equal Option

For the Parametric Test, Variances Assumed Equal option, compute:

$\bar{\bar{X}} = \frac{\sum_{i-1}^{G} N_i \bar{X}_i}{\sum_{i-1}^{G} N_i}$ the overall mean

$BSS = \sum_{i-1}^{G} N_i (\bar{X}_i - \bar{\bar{X}})^2$ the between-groups sum of squares,

$d_1 = G - 1$ the numerator degrees of freedom,

$BMS = BSS / d_1$ the between-groups mean square,

$WSS = \sum_{i-1}^{G} SS_i$ the pooled within-group sum of squares,

$d_2 = \sum_{i-1}^{G} (N_i - 1)$ the denominator degrees of freedom,

if d_2 is not greater than zero, then exit (p-value=1).

$WMS = WSS / d_2$ the within-groups mean square, and

$F = BMS / WMS$ the F-ratio statistic

if $WMS = 0$ then make F is treated as arbitrarily large (p-value = 0).

The p-value is calculated by looking up F in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom.

Computations for Parametric Test, Variances Not Assumed Equal Option

For the Parametric Test, Variances Not Assumed Equal option:

First, check that each group has N_i greater than or equal to 2 and SS_i greater than 0. If not, remove it from consideration and recompute G. If G is not at least 2, exit (p-value=1). (This reflects the more stringent requirements of not assuming the variances equal - if the variance estimate is pooled, replicates are only needed for at least one group, if variances are separately estimated then replicates are needed for each group.)

Then compute:

$$w_i = N_i\left(\frac{N_i - 1}{SS_i}\right) \text{ the group weights}$$

$$W = \sum_{i=1}^{G} w_i \text{ the sum of weights}$$

$$\tilde{X} = \frac{\sum_{i=1}^{G} w_i \overline{X_i}}{W} \text{ the weighted mean}$$

$$BSS = \sum w_i\left(\overline{X_i} - \tilde{X}\right)^2 \text{ the between-groups sum of squares}$$

$$d_1 = G - 1 \text{ the numerator degrees of freedom}$$

$$BMS = BSS/d_1 \text{ the between-groups mean square}$$

$$Z = \frac{1}{G^2 - 1}\sum_{i=1}^{G}\left(1 - \frac{w_i}{W}\right)^2 \Big/ (N_i - 1)$$

$$d_2 = \frac{1}{3Z} \text{ the denominator degrees of freedom}$$

if d_2 is not greater than zero, then exit (p-value=1).

$$WMS = 1 + 2(G - 2)Z \text{ the within-group mean square}$$

$$W = BMS/WMS \text{ the test statistic}$$

The (approximate) p-value is calculated by looking up W in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom. Note that d_2 will not, in general, be an integer.

Computations for Non-parametric Test Option

For the Non-Parametric Test option:

Replace each X_{ik} by R_{ik}, their rank out of all of the $\{X_{ik}\}$ for the gene. Perform the same analysis as for parametric test with variances equal. P-values are approximate but asymptotically accurate.

Multiple Testing Corrections

When testing a set of genes for statistical significance across various groups, some of the genes may be falsely considered as statistically significant. The purpose of a multiple testing correction is to keep the overall error rate/false positives to less than the user-specified p-value cutoff, even if thousands of genes are being analyzed.

For example, if you test 10,000 genes for reliable changes between groups at significance level 0.05, (assuming the tests are independent) you would expect to misidentify about 500 genes as significant, even when there is no real difference in gene expression.

```
10,000 x 0.05 = 500 genes

Possible false positives = (# of genes) (p-value cutoff)
```

If you rely on the nominal p-value when testing the statistical significance of group comparisons for many genes, a significant number of genes pass the filter by chance alone.

Even if you identify 1,000 genes showing significant behavior by this approach, half of the genes on the list appear by chance, which lessens the value of the list. Multiple testing corrections adjust the individual p-value to account for this effect.

Suppose the p-value cutoff is a and the number of genes being tested is N. The first three procedures (Bonferroni, Holm, and Westfall and Young) control the family-wise error rate (FWER), which is the overall probability of obtaining even a single false positive test to be no more than a. This is a very strong criterion, but may be so strong for large lists of genes that no genes are identified as significant. The Benjamini and Hochberg test controls the false discovery rate, defined as the proportion of genes expected to be identified by chance relative to the total number of genes called significant.

Bonferroni

The Bonferroni multiple testing correction, based on Bonferroni's inequality, limits the chance of a false positive results to be no more than a by multiplying each nominal p-value by N (with a maximum of 1). This process controls the FWER, and the expected number of genes by chance is a.

Bonferroni Step Down (Holm)

The Step Down adjustment computes the most significant p-value, and whether it meets the a cutoff after multiplying by *N*. If that gene is found to be significant, the next-most significant gene is considered, but the gene that was found significant is removed from the multiple-testing, so the multiple-testing adjustment is now based on *N* - 1. This process is continued as long as genes pass the successive tests. This process controls the FWER, and expected number of genes by chance is *a*.

Westfall and Young Permutation

This procedure estimates the significance levels of each test by a nonparametric permutation calculation based on the distribution of the significance levels across all possible reassignments of samples to groups. For small numbers of permutations, all permutations are examined.

If there are more than 1000 possible permutations, 1000 of them are selected randomly. P-values are evaluated with respect to this distribution using a step-down procedure as in the Holm procedure. This procedure controls the FWER, and the expected number of genes by chance is *a*.

This test accounts for the dependence structure between genes, and should give a more powerful test than the Bonferroni or Holm procedure. However, the permutation process takes much longer to calculate.

Benjamini and Hochberg False Discovery Rate

In contrast to the Bonferroni, Holm, and Westfall and Young Permutation procedures, the Benjamini and Hochberg procedure controls the False Discovery Rate (FDR). FDR is defined as the proportion of genes expected to occur by chance (assuming genes are independent) relative to the proportion of identified genes.

The expected number of genes by chance is *a* times the number of tests found significant after applying this correction. Since there is no way to calculate this value in advanced, the number of expected genes by chance is 100% times *a.*

This procedure provides a good balance between discovery of significant genes and protection against false positives, since occurrence of the latter is held to a small proportion of the list, and is probably the best choice of multiple testing correction for most situations.

Post Hoc Tests

The one-way ANOVA option determines whether a gene is differentially expressed in any of the conditions tested. It does not, however, indicate which specific group pair(s) are the ones where statistical differences occur. A post hoc test can be used in conjunction with ANOVA to determine which specific group pair(s) are statistically different from each other.

To perform post hoc tests, specify which post hoc test to use in the post hoc test menu. The available choices are Tukey and Student-Newman-Keuls. Pairwise comparisons between all groups are performed. Group comparisons resulting in a p-value below the specified cutoff are displayed in the output. Post hoc tests can only be performed if more than 2 groups are defined by the chosen testing parameter, or if more than 2 groups are chosen manually.

Tukey

In a Tukey test, all means for each condition are ranked in order of magnitude. The group with the lowest mean gets a ranking of 1. The pairwise differences between means, starting with the largest mean compared to the smallest mean, are tabulated between each group pair and divided by the standard error. This value, q, is compared to a Studentized range critical value. If q is larger than the critical value, then the expression between that group pair is considered to be statistically different.

Suppose SGC with more than 2 groups has been performed (ANOVA or Kruskal-Wallis), and that some genes have passed the cutoff. The following calculations are done separately for each gene:

Let $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ be the ordered group means (ascending order). We perform pairwise group mean comparisons in the following manner:

k vs. 1, *k* vs. 2, ... , *k* vs. *k* - 1, then *k* -1 vs. 1, *k* - 1 vs. 2, ... , *k* -1 vs. *k* -2, ending with 2 vs. 1.

Do not perform unnecessary tests. For example, if there is no significant difference between a pair, do not test any “closer” pairs. Each test is performed as described below.

Parametric Test Statistic

If an ANOVA was performed for comparing group means $\bar{X}_a$ and $\bar{X}_b$, compute:

$$SE = \sqrt{\frac{WMS}{n}} \quad \text{if groups are of equal sizes}$$

$$SE = \sqrt{\frac{WMS}{2}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)} \quad \text{if unequal sizes}$$

where n, n_a, and n_b are the corresponding group sizes (number of samples) and WMS is the within-group mean square (from the ANOVA calculations).

Then compute $q = \frac{\bar{X}_b - \bar{X}_a}{SE}$.

Compare this value to the critical value $q_{a,df,k}$ where df is the error degrees of freedom (from the ANOVA calculation) and k is the total number of groups. If q is larger, consider the two means significantly different.

If any of the group sizes for the gene are 1 (for example, there are not replicate samples for every group), do not perform post hoc tests for that gene.

Non-Parametric Test Statistic (Kruskal-Wallis)

If the non-parametric option was chosen, GeneSpring performs a non-parametric Tukey test.

$$SE = \sqrt{\frac{n(nk)(nk+1)}{12}}$$

where n is group size and k is the number of groups.

Rank order all the data and compute rank sums for each group. Order the rank sums and compute q as before, using the rank sums instead of means. Compare q to the critical value $q_{a,\infty,k}$.

Student-Newman-Keuls

The Student-Newman-Keuls test is similar to the Tukey test, except with regard to how the critical value is determined. All a's in Tukey's test are compared to the same critical value determined for that experiment; whereas all a's determined from SNK test are compared to a different critical value. This makes the SNK test slightly less conservative than the Tukey test.

Parametric Test Statistic

When testing group a vs. group b, compare q to $q_{a,df,p}$ where p is the number of means (inclusive) in the range being tested. For example, if comparing group 2 to group 4, $p = 3$.

Non-Parametric Test Statistic

When using the non-parametric test statistic, replace k with p.

Performing Two-way ANOVA

This section explains how to perform a two-way ANOVA. It covers the following topics:

- Performing Two-way ANOVA
- Technical Details for Two-Way ANOVA
- Non-Parametric Two-way ANOVA (Friedman's Test)

Performing Two-way ANOVA

To perform a two-way ANOVA:

1. Select **Tools > Statistical Analysis (ANOVA)**.

   The Statistical Analysis (ANOVA) window opens.

2. In the **Navigator**, select the select the gene list containing the set of genes you would like to analyze, and click the **Choose Gene List button**.

This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and click the **Choose Experiment** button.

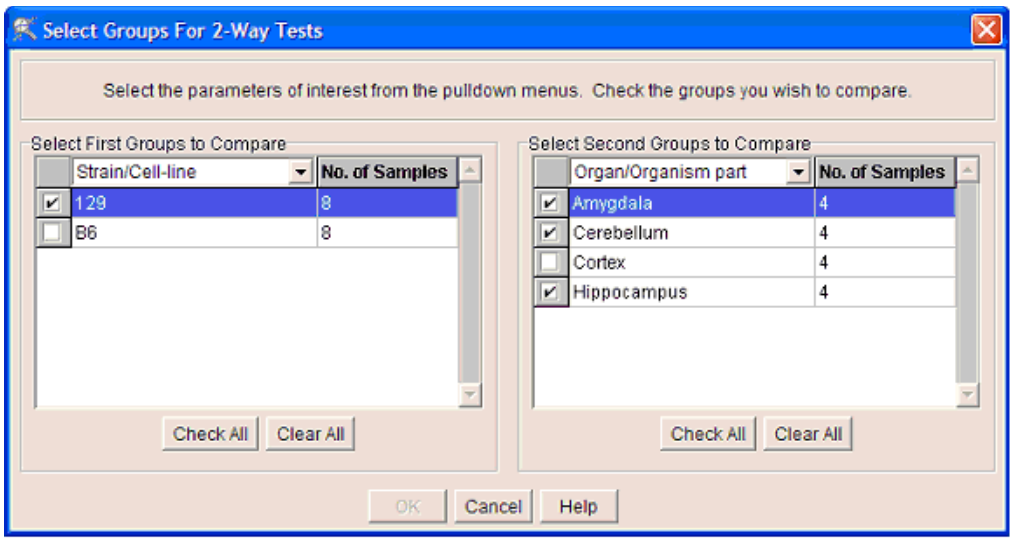
   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. In the **First Parameter to Test** list, do one of the following:
   - Select the parameter on which to base your comparison.
   - To compare only selected conditions for this parameter, go to step 6.

5. In the **Second Parameter to Test** list, do one of the following:
   - Select the parameter on which to base your comparison.
   - To compare only selected conditions for this parameter, go to step 6.

6. Click **Select Groups Manually**.

   The Select Groups for 2-Way Tests window opens.



a. In the Select First Group to Compare or the Select Second Group to Compare table, clear the box in a column's header to ignore groupings based on that parameter.

   Only groups that are checked will be analyzed.

   When you perform this step, the table is dynamically updated to reflect the change. The number of rows decreases and the number of samples associated with each condition increases.

b. Define the conditions to compare by selecting or clearing the check box in a given row.

   The **Check All/Clear All** buttons allow you to select or clear all rows.

7. In the **Test Type** list, select the type of test to perform. There are four testing options:

   - **Parametric test, assume variances equal**—Filter based on the results of a Student's two-sample t-test for two groups or a one-way analysis of variance (ANOVA) for multiple groups.
   - **Parametric test, don't assume variances equal**—Filter based on the results of an ANOVA or Welch's approximate t-test for two groups. This is the most appropriate test for standard experiments in which the global error model is not turned on or should not be used in the analysis.
   - **Parametric test, use all available error estimates**—Filter based on the variances estimated by the Cross-Gene Error Model. If the Cross-Gene Error Model is not turned on, this test is equivalent to the Parametric test.
   - **Non-Parametric test**—Filter based on the rank of each sample, rather than the expression level. Non-parametric comparisons use the Wilcoxon two-sample rank test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups. This test is most successful if you have more than five replicate samples in each group.

8. Select a **P-value cutoff** for genes that pass the filter.

   The p-value indicates a probability, with a value ranging from zero to one, that the difference observed between groups is due to chance. The lower the p-value, the more significant the difference between the groups.

9. Select a type of multiple testing correction.

   Go to "Multiple Testing Corrections" on page 13-12 for a description of the options.

   Go to "Post Hoc Tests" on page 13-13 for a description of the tests.

10. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

    This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

11. Click the **Start** button.

    As the ANOVA executes, you can watch its progress in the Progress bar. Results are shown in the Results window.

12. Go to "Interpreting ANOVA Results" on page 13-23.

Technical Details for Two-Way ANOVA

This section describes the different computations that GeneSpring performs in two-way ANOVA.

Let A and B be the two factors (parameters) chosen by the user. Assume we are looking at a single gene and use the following notation throughout:

Factor *A* has *a* levels, indexed by *i*.

Factor *B* has *b* levels, indexed by *j*.

There are thus *ab* cells of data, each containing at least one value.

Standard Parametric Two-way ANOVA

This test assumes equal variances and equal or proportional replication.

Case 1 - Equal Replication

All cells (groups defined by combinations of the factor levels) have the same number of replicates, say *r*. Therefore, there is a total of *abr* samples.

Let x_{ijk} represent the k^{th} replicate (sample) in level *i* of factor *A* and level *j* of factor *B*.

Let A_i = sum of all observations in level i of factor A

$$= \sum_i \sum_k x_{ijk}$$

Let B_j = sum of all observations in level *j* of factor *B*

$$= \sum_i \sum_k x_{ijk}$$

Let (*AB*)*ij* = sum of all observations in level *i* of factor *A* and level *j* of factor *B*

$$= \sum_k x_{ijk}$$

Let $C = \frac{\left(\sum_i \sum_j \sum_k x_{ijk}\right)}{abr}$

We can now compute the various sums of squares terms:

Total sum of squares:

$$SS(total) = \left(\sum_i \sum_j \sum_k x_{ijk}^2\right) - C$$

Factor A sum of squares:

$$SS(A) = \frac{\sum_i A^2}{rb} - C$$

Factor B sum of squares:

$$SS(B) = \frac{\sum_j B_j^2}{ra} - C$$

Interaction sum of squares:

$$SS(AB) = \frac{\sum_i \sum_j (AB)_j^2}{r} - C - SS(A) - SS(B)$$

Error sum of squares (a.k.a. within-group SS):

$$SS(error) = SS(total) - SS(A) - SS(B) - SS(AB)$$

Now compute the mean sums of squares:

$$MSS(A) = \frac{SS(A)}{a-1}$$

$$MSS(B) = \frac{SS(B)}{b-1}$$

$$MSS(AB) = \frac{SS(AB)}{(a-1)(b-1)}$$

$$MSS(error) = \frac{SS(error)}{abr-ab}$$

Finally, compute F-ratios:

$$F = \frac{MSS(A)}{MSS(error)}$$

$$F = \frac{MSS(B)}{MSS(error)}$$

$$F = \frac{MSS(AB)}{MSS(error)}$$

Each of these should be compared to the upper tail probability of an F distribution with numerator and denominator degrees of freedom given by the corresponding denominators above.

Case II - Proportional Replication

We first check for proportional cell sizes. Let n_{ij} = # of replicates at level i of A and level j of B, and let N be the total number of replicates in all groups = $\sum_i \sum_j n_{ij}$.

If $n_{ij} = \frac{\left(\sum_i n_{ij}\right)\left(\sum_j n_{ij}\right)}{N}$ for each n_{ij}, then we have *proportional replication*.

In this case, all computations are the same as before, with appropriate changes. In particular, the index k in all summations will now go from 1 to n_{ij}, instead of 1 to r.

Let A_i = sum of all observations in level i of factor A

$$= \sum_j \sum_k x_{ijk}$$

Let B_j = sum of all observations in level j of factor B

$$= \sum_i \sum_k x_{ijk}$$

Let $(AB)_{ij}$ = sum of all observations in level i of factor A and level j of factor B

$$= \sum_k x_{ijk}$$

Let $C = \frac{\left(\sum_i \sum_j \sum_k x_{ijk}\right)}{N}$

We can now compute the various sums of squares terms:

Total sum of squares:

$$SS(total) = \left(\sum_i \sum_j \sum_k x_{ijk}^2\right) - C$$

Factor A sum of squares:

$$SS(A) = \sum_i \left(\frac{A_i^2}{\sum_j n_{ij}}\right) - C$$

Factor B sum of squares:

$$SS(B) = \sum_j \left(\frac{B_j^2}{\sum_i n_{ij}}\right) - C$$

Interaction sum of squares:

$$SS(AB) = \sum_i \sum_j \left(\frac{(AB)_{ij}^2}{n_{ij}} \right) - C - SS(A) - SS(B)$$

$$SS(error) = SS(total) - SS(A) - SS(B) - SS(AB)$$

Compute mean sums of squares using the following degrees of freedom:

Factor *A* df = *a* - 1

Factor *B* df = *b* - 1

Interaction (*AB*) df = (*a* - 1)(*b* - 1)

Error df = *N* - *ab*

Divide each SS by appropriate degrees of freedom to obtain mean SS. F-ratios and p-values are then computed as before.

Case III - No Replicates (One Per Cell)

We can still perform Two-way ANOVA in this case, but it is not possible to test for an interaction effect. The calculations are essentially the same as before.

The total number of replicates is $N = ab$.

$$C = \frac{\left(\sum_i \sum_j x_{ij} \right)}{N}$$

$$SS(total) = \sum_i \sum_i x_{ij}^2 - C$$

$$SS(A) = \frac{\sum_i \left(\sum_j x_{ij} \right)}{b} - C \quad \text{degrees of freedom} = a - 1$$

$$SS(B) = \frac{\sum_j \left(\sum_i x_{ij} \right)}{a} - C \quad \text{degrees of freedom} = b - 1$$

$$SS(error) = SS(total) - SS(A) - SS(B) \quad \text{degrees of freedom} = (a - 1)(b - 1)$$

Mean sums of squares = sum of squares / degrees of freedom.

F ratios:

$$F = \frac{MSS(A)}{MSS(error)}$$

$$F = \frac{MSS(B)}{MSS(error)}$$

Compute p-values as before.

Case IV - Disproportional Replication

If a single cell is one value short of the number required for proportional replication, estimate the missing value using the following:

$$x_{ijk} = \frac{aA_i + bB_j - \sum_i \sum_j \sum_k x_{ijk}}{N + 1 - a - b}$$

where A_i,B_j are as before and N is the total number of data *including the missing value.*

If several cells are missing values, or more than one value is missing, apply this formula iteratively if necessary.

After missing values have been estimated, perform ANOVA calculations as above, but do not increase degrees of freedom. That is, error df should still be based on the original number of data points.

GeneSpring displays a warning message if missing values have been imported.

If there is a larger number of missing values (say >min(a,b)), GeneSpring displays a warning and exits.

Non-Parametric Two-way ANOVA (Friedman's Test)

The Friedman's Test only tests for the effect of factor A, while controlling for factor B. To get a p-value for the other factor, reverse the factors (parameters). This does *not* test for interaction.

Case I - No Replicates (One Sample per Cell)

Rank data within each of the *b* blocks separately. For each of the *a* levels of factor A, compute rank sums R_i.

Then compute

$$\chi_r^2 = \frac{12}{ba(a+1)} \sum_{i=1}^{a} R_i^2 - 3b(a+1)$$

This statistic has its own distribution; however, we can approximate by computing:

$$F_F = \frac{(b-1)\chi_r^2}{b(a-1) - \chi_r^2}$$

and compare this to *F* with degrees of freedom *a* - 1 and (*a* - 1)(*b* - 1).

Case II - Equal Replicates (One per Cell

If there are *n* replicates within each cell, compute

$$\chi_r^2 = \frac{12}{ban^2(na+1)} \sum_{i=1}^{a} R_i^2 - 3b(na+1)$$

Compare this to the chi-square critical value with *a* - 1 degrees of freedom.

Interpreting ANOVA Results

This section explains how to interpret the test results for one-way ANOVA and two-way ANOVA.

ANOVA without Post Hoc Test

After one-way ANOVA has been performed, the New Gene List window opens. The Notes section indicates what setting was used for this analysis and the percentage of genes that could have been identified by chance. If the genes in this gene list were found to have measurements considered statistically different across at least one group-pair, you would not be able to determine which group was differentially expressed from this analysis. You need to run a post hoc test to determine which group was different.

The gene list will have an associated value that represents the p value as calculated for the ANOVA. This associated p-value can be used in further filtering. Go to"Viewing Generated P-values" on page 13-27 for more information. The associated p-values can also be plotted on the scatter plots. Go to "Using the Scatter Plot View" on page 8-20 for more information.

ANOVA with Post Hoc Tests

Summary by Gene

The Results Summary by Gene tab displays the mean expression level by group for each significant gene. For each gene, the coloring indicates which groups differ significantly from the others. Groups of the same color show no significant difference. Groups of different colors differ significantly from each other.

For example, Figure 13-2 lists all the genes considered differentially expressed by statistical criteria. Groups with the highest color differential have the most significant difference. Groups with the same color show no statistical difference for that gene. A group colored grey is considered to be unknown because the significance of its mean difference cannot be determined with confidence from the test used.

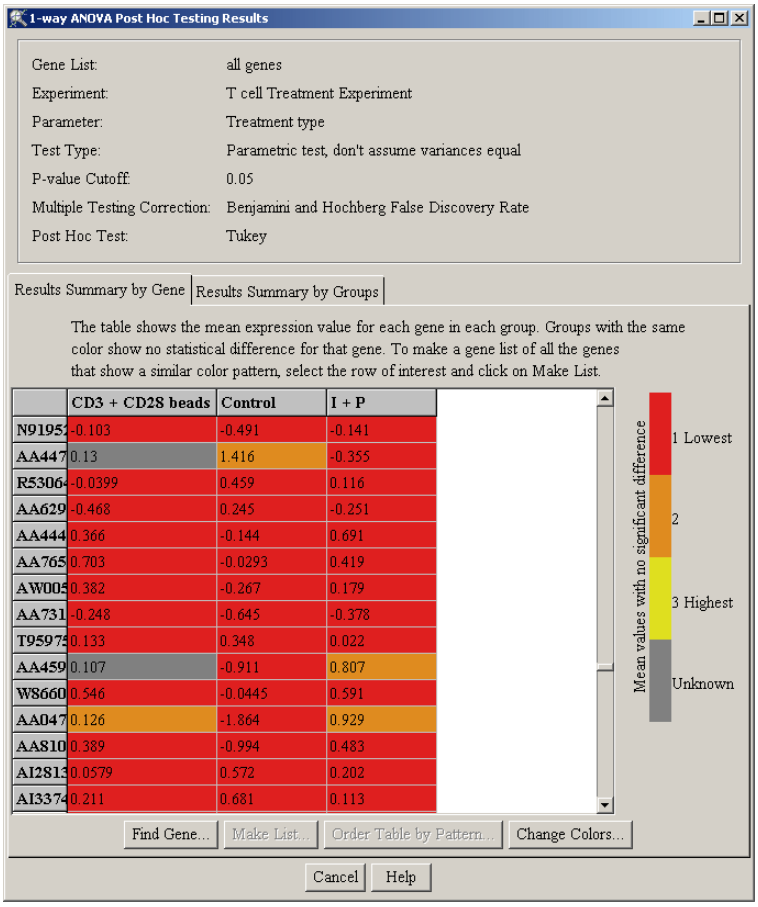

1-way ANOVA Post Hoc Testing Results
Gene List: all genes
Experiment: T cell Treatment Experiment
Parameter: Treatment type
Test Type: Parametric test, don't assume variances equal
P-value Cutoff: 0.05
Multiple Testing Correction: Benjamini and Hochberg False Discovery Rate
Post Hoc Test: Tukey
Results Summary by Gene
Results Summary by Groups
The table shows the mean expression value for each gene in each group. Groups with the same color show no statistical difference for that gene. To make a gene list of all the genes that show a similar color pattern, select the row of interest and click on Make List.
CD3 + CD28 beads
Control
I + P
N91951 -0.103 -0.491 -0.141
AA447 0.13 1.416 -0.355
R53064 -0.0399 0.459 0.116
AA629 -0.468 0.245 -0.251
AA444 0.366 -0.144 0.691
AA765 0.703 -0.0293 0.419
AW005 0.382 -0.267 0.179
AA731 -0.248 -0.645 -0.378
T95975 0.133 0.348 0.022
AA459 0.107 -0.911 0.807
W8660 0.546 -0.0445 0.591
AA047 0.126 -1.864 0.929
AA810 0.389 -0.994 0.483
AI2813 0.0579 0.572 0.202
AI3374 0.211 0.681 0.113
Mean values with no significant difference
1 Lowest
2
3 Highest
Unknown
Find Gene...
Make List...
Order Table by Pattern...
Change Colors...
Cancel
Help

Figure 13-2 1-way ANOVA with Post Hoc Test, Summary by Gene Tab

Summary by Groups

The Results Summary by Groups tab displays a matrix with rows and columns indexed by parameter values. Each cell corresponds to a combination of groups. The numbers in the lower half of the matrix represent the number of genes that differ significantly between the groups. The numbers in the upper half are the genes which show no significant difference.

For example, Figure 13-3 indicates the total number of genes that are statistically differentially expressed between the groups being compared in the matrix. Greater color saturation indicates greater difference (or similarity). Total number of genes analyzed is shown in the box colored grey. Gene lists can be generated from each, or combination of the boxes, by highlighting the appropriate boxes and selecting Make List of Union or Make List of Intersection.

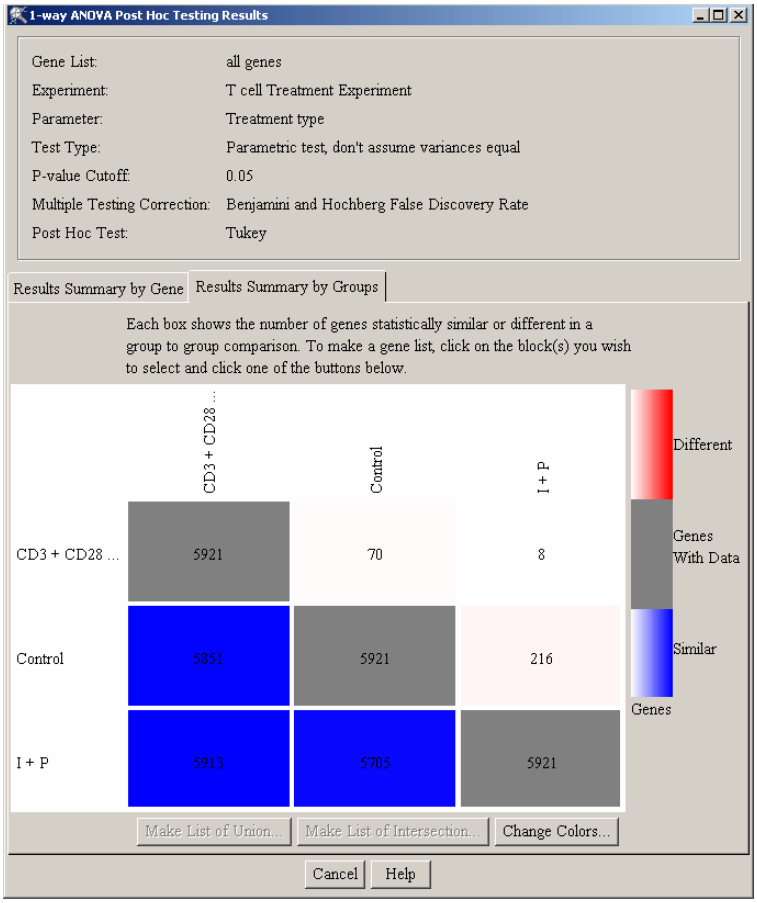
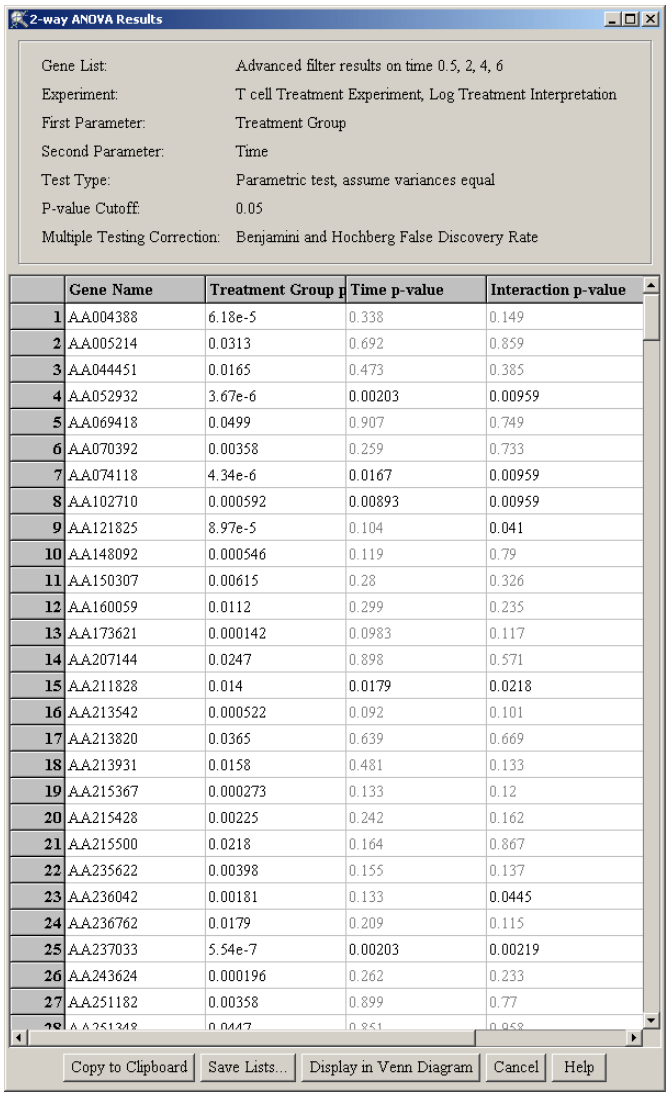


Figure 13-3 1-way ANOVA with Post Hoc Test, Summary by Groups Tab

Two-way ANOVA

Results for a two-way ANOVA appear in the window (Figure 13-4). This section describes the elements of the window.


2-way ANOVA Results
Gene List: Advanced filter results on time 0.5, 2, 4, 6
Experiment: T cell Treatment Experiment, Log Treatment Interpretation
First Parameter: Treatment Group
Second Parameter: Time
Test Type: Parametric test, assume variances equal
P-value Cutoff: 0.05
Multiple Testing Correction: Benjamini and Hochberg False Discovery Rate
Copy to Clipboard
Save Lists...
Display in Venn Diagram
Cancel
Help

	Gene Name	Treatment Group p	Time p-value	Interaction p-value
1	AA004388	6.18e-5	0.338	0.149
2	AA005214	0.0313	0.692	0.859
3	AA044451	0.0165	0.473	0.385
4	AA052932	3.67e-6	0.00203	0.00959
5	AA069418	0.0499	0.907	0.749
6	AA070392	0.00358	0.259	0.733
7	AA074118	4.34e-6	0.0167	0.00959
8	AA102710	0.000592	0.00893	0.00959
9	AA121825	8.97e-5	0.104	0.041
10	AA148092	0.000546	0.119	0.79
11	AA150307	0.00615	0.28	0.326
12	AA160059	0.0112	0.299	0.235
13	AA173621	0.000142	0.0983	0.117
14	AA207144	0.0247	0.898	0.571
15	AA211828	0.014	0.0179	0.0218
16	AA213542	0.000522	0.092	0.101
17	AA213820	0.0365	0.639	0.669
18	AA213931	0.0158	0.481	0.133
19	AA215367	0.000273	0.133	0.12
20	AA215428	0.00225	0.242	0.162
21	AA215500	0.0218	0.164	0.867
22	AA235622	0.00398	0.155	0.137
23	AA236042	0.00181	0.133	0.0445
24	AA236762	0.0179	0.209	0.115
25	AA237033	5.54e-7	0.00203	0.00219
26	AA243624	0.000196	0.262	0.233
27	AA251182	0.00358	0.899	0.77

Figure 13-4 2-way ANOVA Results Window

Analysis Information

The top of the window displays the gene list name, experiment, and parameters used, as well as the test type, p-value cutoff, and multiple testing correction that was performed.

Table

- **Gene Name**—Displays the name of the gene in the experiment.
- **First Parameter**—Displays the name of the first parameter you selected.
- **Second Parameter**—Displays the name of the second parameter you selected.
- **P-value**—Indicates a probability, with a value ranging from zero to one, that the difference observed is due to a coincidence of random sampling. If the p-value is small, then it is unlikely that the difference observed is due to a coincidence of random

sampling. You can reject the hypothesis that the populations have identical means. If the p-value is large, the results do not provide sufficient evidence to conclude that the means of the groups differ.

- **Interaction P-value**—Indicates a probability, with a value ranging from zero to one, that the interaction effects between groups is due to a coincidence of random sampling. If the p-value is small, then it is unlikely that the interaction effects between groups is due to a coincidence of random sampling. You can reject the hypothesis that the populations have identical means. If the p-value is large, the results do not provide sufficient evidence to conclude that the means of the groups differ.

From this window, you can do the following:

- **Copy to Clipboard**—Copy the results in this window to the clipboard. You can paste these results into a spreadsheet program or text editor.
- **Save Lists**—Save the results as a gene list or lists.
- **Display in Venn Diagram**—Display the results in a Venn diagram in the main GeneSpring window.

Viewing Generated P-values

The associated p-value for the genes on this gene list can be viewed in GeneSpring using the following methods:

- **Gene List Inspector**—Double-click on the selected gene list to open up the Gene List Inspector window. P-values are shown under the p-value columns.
- **Ordered List**—Select the gene list and go to **View > Ordered List**. Genes are displayed according to p-values: smallest p-values are on the left-hand side, highest p-values are on the right-hand side.
- **Export Annotated Gene List**—Highlight the gene list and go to **Edit > Copy > Copy Annotated Gene List** and select to export out Gene List Associated Values.

References

Benjamini, Y. and Hochberg, Y. (1995) “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society B*, 57, 289 -300.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2000) “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”. *Department of Statistics Technical Report* #578, University of California, Berkeley (http://stat-ftp.berkeley.edu/tech-reports/index.html)

Holm, S. (1979) “A Simple Sequentially Rejective Bonferroni Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65 -70.

Miller, R.G. (1981) *Simultaneous Statistical Inference*, Second Edition. New York: Springer-Verlag.

Westfall, P.H. and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*. New York: John Wiley & Sons, Inc.

14

Finding Genes with Similar Expression Profiles

This chapter explains how to use correlation techniques to discover new gene regulatory elements or find similar expression patterns. This chapter covers the following topics:

- Finding Potential Regulatory Sequences
- Finding Similar Genes
- Correlation of Expression Change

Chapter 15, "Clustering and Characterizing Data" explains how to find similar expression patterns using clustering and other data characterization techniques.

Finding Potential Regulatory Sequences

The Find Potential Regulatory Sequence feature is useful for finding genes sharing similar regulatory sequences or having a particular regulatory sequence in common. By correlating gene expressions with promoter sequences, you can discover new gene regulatory elements. You need a whole genomic sequence of the organism to use this analysis, or at least the upstream region for the genes in your genome.

Once you have found the regulatory sequences you want, you can create gene lists from them.

This section contains the following topics:

- Find Potential Regulatory Sequence Window
- Finding Regulatory Sequences
- Entering Specific Regulatory Sequences
- Viewing Regulatory Sequence Search Results
- Interpreting Regulatory Sequence Search Results
- Viewing the Conjectured Nucleotide Sequence

Find Potential Regulatory Sequence Window

The Find Potential Regulatory Sequence window (Figure 14-1) lets you find common regulatory sequences upstream of genes in a gene list, or to search for a known sequence. It also compares the frequency of occurrence against all other gene lists in the genome.

When the regulatory sequences tool compares the upstream region of genes to the remainder of the genome, it uses the “all genes” list. The “all genomic elements” list includes non-gene elements that are not expressed.

Note: In GeneSpring version 4.0 and later, the sequence information is loaded automatically. To change the load automatically feature, select **Edit > Preferences > Data Files** and clear the Load Sequence check box.

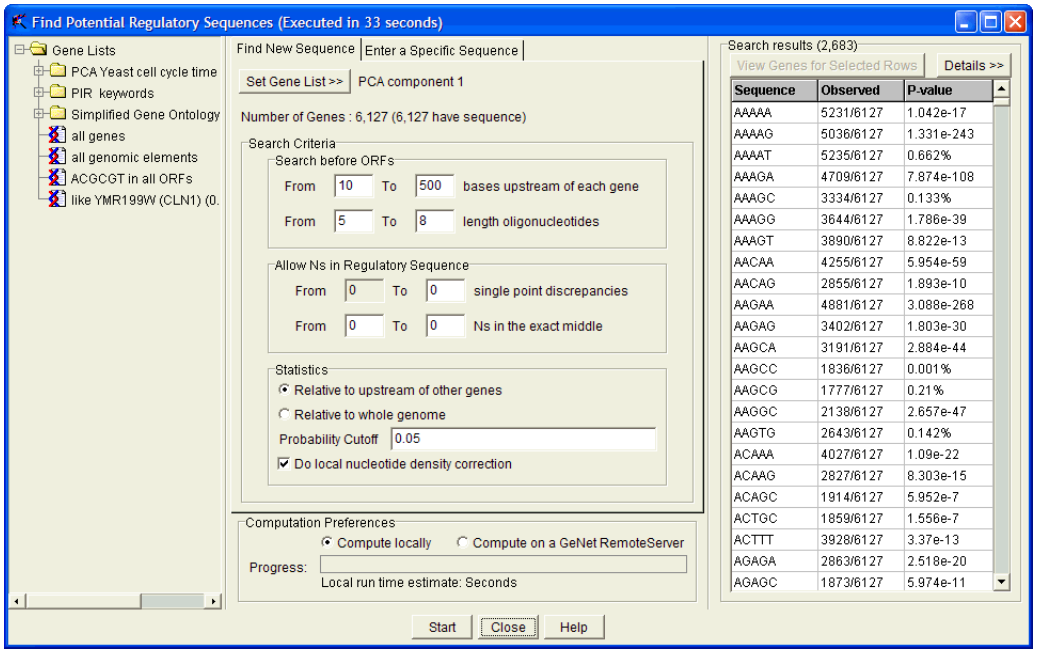


Figure 14-1 Find Potential Regulatory Sequences Window

From this window you can search for short sequences upstream of the genes in the current gene list or across the entire genome. You can also enter a known sequence to examine.

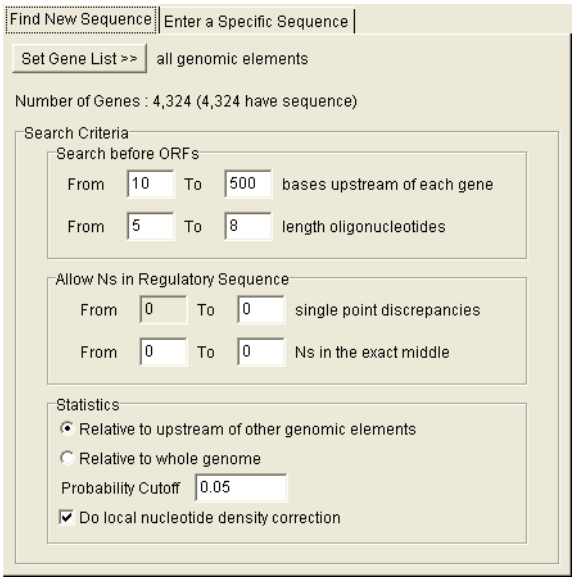
Finding Regulatory Sequences

To find a new regulatory sequence:

1. Select **Tools > Find Potential Regulatory Sequences**.

   The Find Potential Regulatory Sequences window opens.

2. Click the **Find New Sequence** tab.


Find New Sequence
Enter a Specific Sequence
Set Gene List >>
all genomic elements
Number of Genes : 4,324 (4,324 have sequence)
Search Criteria
Search before ORFs
From 10 To 500 bases upstream of each gene
From 5 To 8 length oligonucleotides
Allow Ns in Regulatory Sequence
From 0 To 0 single point discrepancies
From 0 To 0 Ns in the exact middle
Statistics
Relative to upstream of other genomic elements
Relative to whole genome
Probability Cutoff 0.05
Do local nucleotide density correction

3. In the **Navigator**, select the gene list you want and click **Set Gene List**.

   Note: Do not choose the **All Genes** or **All Genomic Elements** gene lists. You are already comparing your selected gene list against all other genes in the genome.

4. Enter the number of bases upstream of each gene in the **Search Before ORFs** section of the window.

   For example, if you enter “From 10 To 100” on a search for ACGCGT, GeneSpring searches for any part of the promoter within the region between 10 and 100. The smaller the range between these numbers, the greater the likelihood that the results are statistically significant.

   Larger sequences may take longer to search. You can also search for common sequences within the ORF by using negative numbers for the bases.

5. Enter the length of the oligonucleotides you want to search.

6. Enter the number of single point discrepancies allowed.

   This option refers to a maximum number of mismatches allowed. For example, if you specify one single point discrepancy, ACGCG**A**T satisfies a search for ACGCG**T**T.

7. Enter the range of base gaps in the exact middle.

   This option refers to the size of an allowable hole in the middle of the sequence, allowing you to look for sequences such as ACGnnnCGT, which is biologically relevant due to loops and non-binding areas.

   The gap must be in the exact middle, with the longer side of odd sequences appearing before the Ns. It does not count towards the sequence length specified; hence ACGnnnCGT would be returned as an oligonucleotide of length 6.

8. Select whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence.

   Note: The first option is far more common.

The Probability Cutoff text box indicates the level of significance (P-value) needed for an oligomer to be listed in the results. You can change this value.

9. Specify whether to perform the operation locally or on a remote Signet server.
10. Click **Start**.
11. The progress bar depicts the length of the operation. For very large genomes or complex search parameters, this operation may take a few minutes.

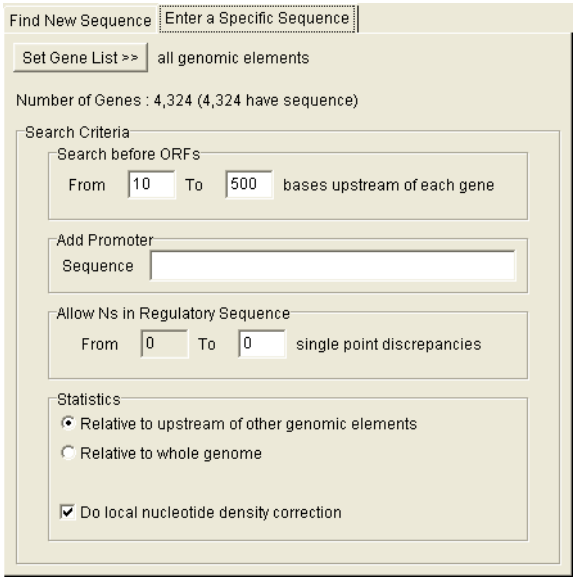
Entering Specific Regulatory Sequences

To enter a specific regulatory sequence:

1. Select **Tools > Find Potential Regulatory Sequences**.

   The Find Potential Regulatory Sequences window opens.

2. Click the **Enter a Specific Sequence** tab.



3. In the **Navigator**, select the gene list you want and click **Set Gene List**.

   Note: Do not choose the **All Genes** or **All Genomic Elements** gene lists. You are already comparing your selected gene list against all other genes in the genome.

4. Enter the number of bases upstream of each gene in the **Search Before ORFs** section of the window.

   For example, if you enter “From 10 To 100” on a search for ACGCGT, GeneSpring searches for any part of the promoter within the region between 10 and 100. The smaller the range between these numbers, the greater the likelihood that the results are statistically significant.

   Larger sequences may take longer to search. You can also search for common sequences within the ORF by using negative numbers for the bases.

5. Enter the promoter sequence in the **Sequence** box.

The sequence can be entered using the following the IUPAC codes.

Table 14-1 Summary of Single-letter IUPAC Codes

Symbol	Meaning	Origin of Designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

6. Enter the number of single point discrepancies allowed.

   This option refers to a maximum number of mismatches allowed. For example, if you specify one single point discrepancy, ACGCG**A**T satisfies a search for ACGCG**T**T.

7. Select whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence.

   Note: The first option is far more common.

   The Probability Cutoff text box indicates the level of significance (P-value) needed for an oligomer to be listed in the results. You can change this value.

8. Specify whether to perform the operation locally or on a remote Signet server.

9. Click **Start**.

   The progress bar depicts the length of the operation. For very large genomes or complex search parameters, this operation may take a few minutes.

Viewing Regulatory Sequence Search Results

The search results are displayed in the Results area of the Find Potential Regulatory Sequences window.

To view regulatory sequence search results:

- Do one of the following:
  - Click **Details** for expanded results data.
  - Click **View Genes for Selected Row**.

Interpreting Regulatory Sequence Search Results

The Results area of the Find Potential Regulatory Sequences window provides the following information:

- **Sequence—**The nucleotide sequence of the oligomer.
- **Observed—**The number of genes in the list where the oligomer was found.
- **P-value—**The probability (p-value) that the number of occurrences in the list came about by chance. Only nucleotide motifs with P-values below the specified probability cutoff (in this case 0.05 or 5%) are shown.
- **Random Rate—**The intrinsic probability, which is the percent of genes for which you would expect this specific nucleotide combination to appear upstream (if the nucleotide sequence was strictly random). Although the nucleotide sequence is not random, the rate is still a good value that can be compared to the observed probability.
- **Observed: Other Genes—**The observed probability of this sequence motif appearing upstream of genes other than the list under inspection. If the option **Relative to sequence upstream of other genes** is selected, this becomes the probability of the observed sequence occurring relative to the genes not in the list; for example, relative to the "all genes" list.

  If the option **Relative to whole genomic sequence** is selected, this becomes the probability of one or more occurrences of the sequence based on the rate of occurrence in the entire genome.

  The formula used to calculate this is:

  `1-(1-k/b)n`

  where:

  - `k` = the number of occurrences in the whole sequence
  - `b` = the total number of bases
  - `n` = the length of the upstream region being searched
- **Expected—**The number of incidences in the searched gene list in which you would expect this oligomer to occur. The number for the *Expected* column is derived using the larger of the intrinsic probability and the observed probability values.

- **Single P—**This column displays the Single P value for the motif. If only one test was performed, this value indicates the probability that this particular sequence would be found by chance.
- **Tests—**The number of tests run to generate these motifs appears in the last column. This is the number of oligomers tested that were the length of the sequence motif found.

Viewing the Conjectured Nucleotide Sequence

After finding the common regulatory sequence, you can also view the conjectured regulatory sequence.

Conjectured Regulatory Sequence window

The Conjectured Regulatory Sequence window (Figure 14-2) displays the common nucleotide sequence, showing the 10 bases that precede and follow it in the area near (or in) each gene where the oligomer is found.

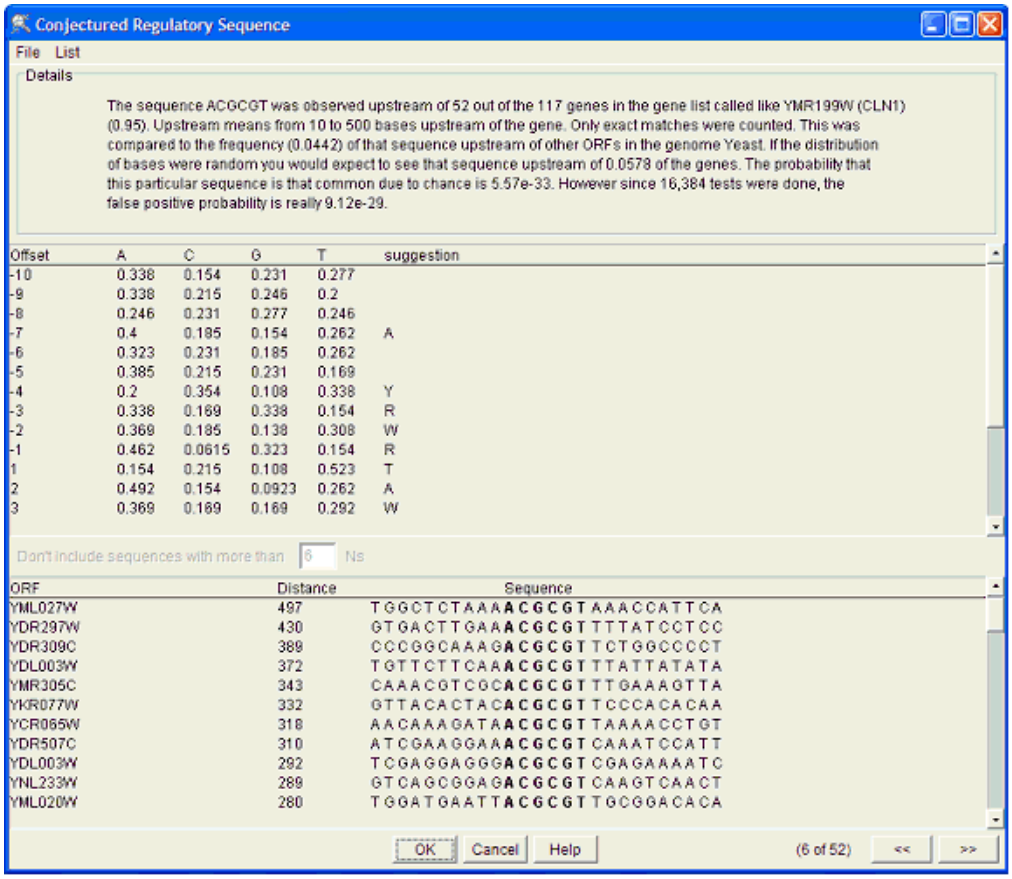

Conjectured Regulatory Sequence
File List
Details
The sequence ACGCGT was observed upstream of 52 out of the 117 genes in the gene list called like YMR199W (CLN1) (0.95). Upstream means from 10 to 500 bases upstream of the gene. Only exact matches were counted. This was compared to the frequency (0.0442) of that sequence upstream of other ORFs in the genome Yeast. If the distribution of bases were random you would expect to see that sequence upstream of 0.0578 of the genes. The probability that this particular sequence is that common due to chance is 5.57e-33. However since 16,384 tests were done, the false positive probability is really 9.12e-29.
Offset A C G T suggestion
-10 0.338 0.154 0.231 0.277
-9 0.338 0.215 0.246 0.2
-8 0.246 0.231 0.277 0.246
-7 0.4 0.185 0.154 0.262 A
-6 0.323 0.231 0.185 0.262
-5 0.385 0.215 0.231 0.169
-4 0.2 0.354 0.108 0.338 Y
-3 0.338 0.169 0.338 0.154 R
-2 0.369 0.185 0.138 0.308 W
-1 0.462 0.0615 0.323 0.154 R
1 0.154 0.215 0.108 0.523 T
2 0.492 0.154 0.0923 0.262 A
3 0.369 0.169 0.169 0.292 W
Don't include sequences with more than 6 Ns
ORF Distance Sequence
YML027W 497 TGGCTCTAAAACGCGTAAACCATTCA
YDR297W 430 GTGACTTGAAACGCGTTTTATCCTCC
YDR309C 389 CCCGGCAAAGACGCGTTCTGGCCCCT
YDL003W 372 TGTTCTTCAAACGCGTTTATTATATA
YMR305C 343 CAAACGTCGCACGCGTTTGAAAGTTA
YKR077W 332 GTTACACTACACGCGTTCCCACACAA
YCR065W 318 AACAAAGATAACGCGTTAAAACCTGT
YDR507C 310 ATCGAAGGAAACGCGTCAAATCCATT
YDL003W 292 TCGAGGAGGGACGCGTCGAGAAAATC
YNL233W 289 GTCAGCGGAGACGCGTCAAGTCAACT
YML020W 280 TGGATGAATTACGCGTTGCGGACACA
OK Cancel Help (6 of 52) << >>

Figure 14-2 Conjectured Regulatory Sequence Window

The following section describes the elements in this window.

Sections

- **Details—**Provides a general description of the common sequence motif being inspected. The details found in this box are the same numbers listed in the right-hand columns of the Results box in the Find Potential Regulatory Sequences window.
- **Offset Bases—**The middle third of the Conjectured Regulatory Sequence window contains statistics on the bases to either side of the motif. The first column contains the offset from the observed sequence. The next four columns contain the percentage of genes with that base in that position. The last column contains a suggested extension to the motif.
- **ORF—**The bottom third of the Conjectured Regulatory Sequence window contains the sequence information for the motif being inspected, as it occurs in the nucleotide sequence in the area near (or in) each gene where it is found. There are three columns of data.
  - **ORF—**Indicates the gene for which the common sequence motif (given in bold, centered in the column) is upstream.
  - **Distance—**Displays the number of bases upstream that the oligomer is from the ORF associated with it in the first column. This number is the difference between the base pair number of the first base in the gene and the base pair number of the first nucleotide in the motif. It includes the distance of the promoter. This means the distance number is the difference between the promoter sequence and the ORF.
  - **Sequence—**Contains the sequence being examined, displayed in bold. On the left side of it are the ten bases proceeding this instance of the motif, and on the right side are the 10 bases that follow it in the nucleotide sequence.

This window also provides a brief description of the statistics listed in the Results box of the Find Potential Regulatory Sequences window, and lets you modify the observed motif by removing an item, extending the promoter or making a new gene list.

File Menu

The File menu contains the following commands:

- **Print—**Prints the list in the lower half of the Conjectured Regulatory Sequence window.
- **Close—**Closes the Conjectured Regulatory Sequence window

List Menu

The List menu contains the following commands:

- **Remove Item—**Removes the highlighted item and its associated sequence motif from the list matching the common sequence motif being examined.
- **Make Gene List—**Displays the new Gene List window.

  When a gene list is based on the occurrence of a specified sequence, GeneSpring associates a number with each gene corresponding to the distance of the first sequence that is upstream of the ORF. The numbering begins from the first nucleotide sequence.

Zoom in on the Ordered list view or open the Gene List Inspector to view these numbers.

- **Extend Promoter—**Adds a new and longer promoter in the Find Potential Regulatory Sequences window.

Viewing the Conjectured Nucleotide Sequence

To view the conjectured nucleotide sequence:

- Double-click one of the sequence motifs given in the **Results** box.

Making Gene Lists from Conjectured Nucleotide Sequences

Once you have found possible regulatory sequences using the Find Potential Regulatory Sequences window and are inspecting one of the sequences in the Conjectured Regulatory Sequence window, you can make a list of all of the genes containing that sequence by selecting **List > Make Gene List**.

Finding Similar Genes

If you have a single gene that shows an interesting expression pattern, you can find genes that show a similar expression pattern by using the Find Similar Genes feature.

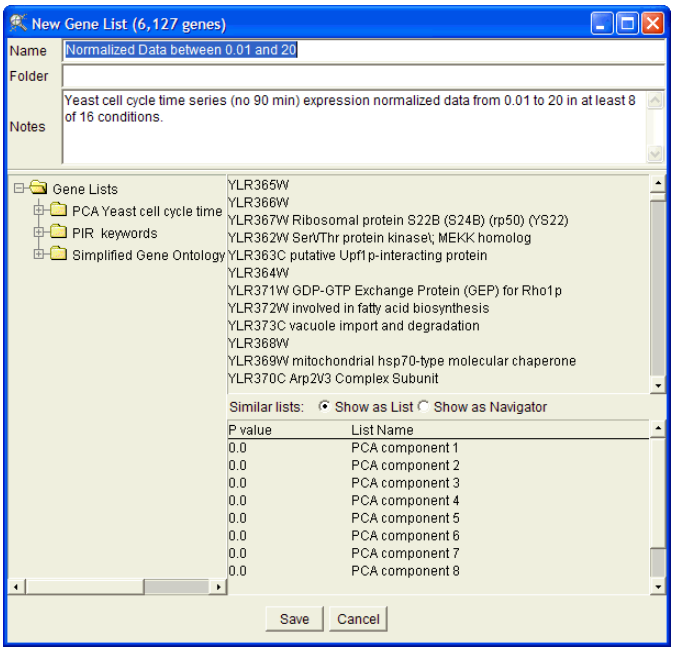
To find similar genes:

1. Do one of the following:
   - Click on a gene in the browser window and select **Edit > Inspect Selected Gene**
   - Double-click a gene (this may be easier when you zoom in).
   - Select **Edit > Find Gene** and enter the name of the gene.
   - Press **Ctrl+I** (when one or more genes are selected).

   The Gene Inspector window opens.

   Go to “Using the Gene Inspector” on page 7-16 for more information.
2. Click the **Find Similar** button.

   The New Gene List window opens.



This window includes the genes in that list, as well as lists that are similar to a gene list.

3. Enter a name for the new gene list or accept the default.
4. To assign the gene list to a project, click the **Change Projects** button.

   The Change Projects window opens.
5. To create a new project name, do the following:
   a. Click the **Add New** button.

      The Add New window opens.
   b. In the **Project Name** field, enter a name for the project.

      All data objects in the project will be associated with this project name.
   c. Click the **OK** button.
6. To assign a previously created project name, do the following:
   a. Select the project name(s) you want.

      You can assign data objects to more than one project.
   b. Click the **OK** button.
7. Click the **Save** button.

Correlation of Expression Change

Correlation is a measure of the relationship between two or more variables. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents no correlation

GeneSpring enables you to a search for genes with similar expression profiles using the following correlation tests:

- Standard Correlation
- Smooth Correlation
- Change Correlation
- Upregulated Correlation
- Pearson Correlation
- Spearman Correlation
- Spearman Confidence
- Two Sided Spearman Confidence
- Distance

Complex Correlations Window

You can use the Complex Correlations window (Figure 14-3) to set up complex correlations against the inspected gene. These correlations can involve more than one experiment or condition or extra restrictions on experiments.

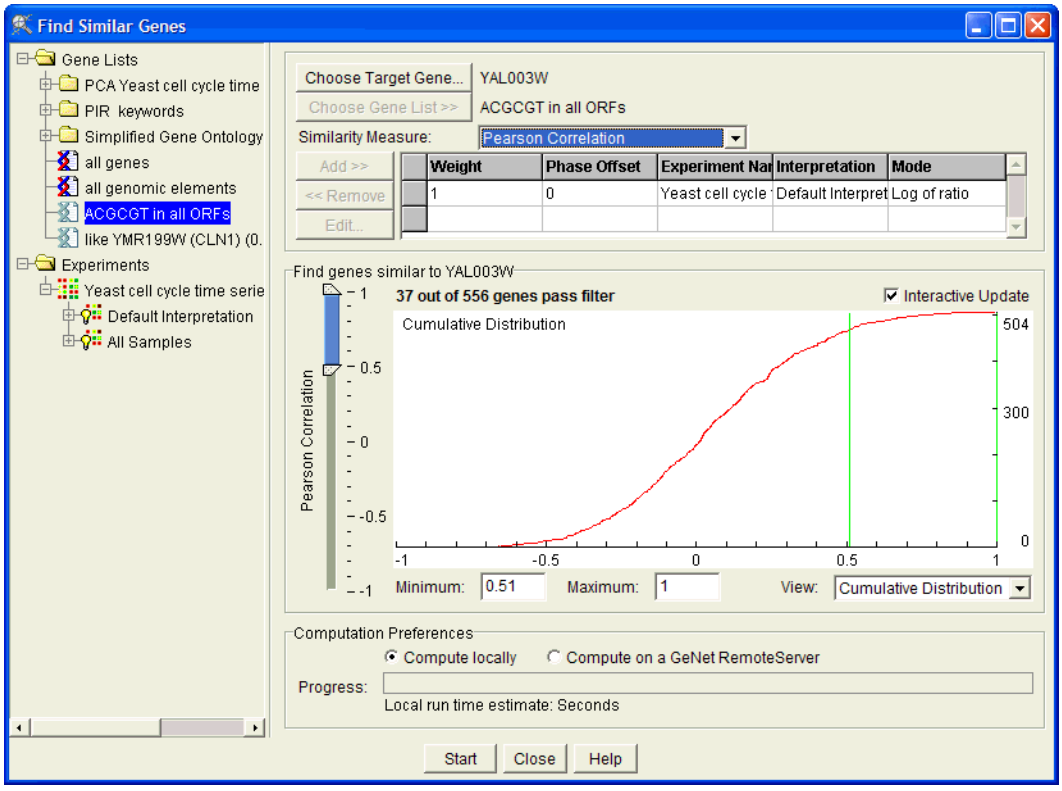


Figure 14-3 Find Similar Genes Window: Complex Correlations

Preview Pane

You can view the Preview pane as a cumulative distribution, a graph, or by linking the preview to the main GeneSpring window. When you are working with large experiments, you may want to clear the Interactive Update box to the upper right of the Preview pane to avoid slowing the analysis.

Graph Panel

A cumulative distribution graph of gene correlations appears in the center of the New Correlation window. The horizontal axis shows the correlation from zero to 1. The vertical axis represents the number of genes. The green lines are your specified maximum and minimum values. If you change these values the green lines move accordingly.

Correlations Table

The Correlations table lists the experiments chosen to correlate against the specified gene. The experiments selected may be weighted, making one more important than another. If both experiments chosen are given a weight of 1, they are averaged equally. To modify an experiment's weight, click in the Weight column to the left of its name and enter a new weight.

When you are interested in genes that show a similar expression profile, but slightly offset in phase you can use the phase offset feature to find those genes. This is particularly useful in those cases where the experiment is a time series, where you want to find those genes that are regulated by other genes. Genes that are regulated by certain genes usually show the same expression profile, but with a slight delay and using the phase offset feature allows you to find those genes.

The phase offset can only be used when the conditions used in the graph is based on one or more numerical parameters. The number you enter in the Phase Offset window is in the same units as the original parameter.

Equation

The equation used to determine the overall correlation is:

$$\mathbf{X = \frac{(Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}}$$

Variables

The variables in the equation are:

- **A—**The correlation coefficient between the gene in question in experiment 1 and the selected gene, also from experiment 1.
- **a—**The weight specified for experiment 1.
- **B—**The correlation coefficient of the gene in question in experiment 2, to the selected gene, also from experiment 2.
- **b—**b is the weight associated with experiment 2.
- **C—**The correlation coefficient of the gene in question in experiment 3 to the selected gene, also from experiment 3.
- **c—**The weight associated with experiment 3.

Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white Correlations table. If **X** is between the minimum and maximum correlations specified in the Find Similar Genes window, the gene in question passes the correlations.

Table 14-2 describes the different correlations that are available in the Find Similar Genes window.

Table 14-2 Find Similar Genes Window—Correlations

Standard Correlation	Measures the angular separation of expression vectors for Genes **A** and **B** around zero. *Result* = **a.b**/(|**a**||**b**|)
Smooth Correlation	Make a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph window. Do the same to make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)
Change Correlation	Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is atan(a_{i+1}/a_i)-π/4.Do the same to make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)
Upregulated Correlation	Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is max(atan(a_{i+1}/a_i)-π/4,0). Do the same to make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)
Pearson Correlation	Calculate the mean of all elements in vector **a**. Then subtract that value from each element in **a**. Call the resulting vector **A**. Do the same for **b** to make a vector **B**. *Result* = **A.B**/(|**A**||**B**|)
Distance	Distance is not a correlation at all, but a measurement of dissimilarity. Distance is the measurement of Euclidian distance between the expression profile for gene **A** (defined by its expression values for each point in N-dimensional space, where N is the number of conditions with data in your experiment) and the expression profile for gene **B**. *Result* = |**a-b**| divided by the square root of the number of conditions with data
Spearman Correlation	Order all the elements of vector **a**. Use this order to assign a rank to each element of **a**. Make a new vector **a**' where the i^{th} element in **a**' is the rank of a_i in **a**. Now make a vector **A** from **a**' in the same way as **A** was made from **a** in the Pearson Correlation. Similarly, make a vector **B** from **b**. *Result* = **A.B**/(|**A**||**B**|)

Table 14-2 Find Similar Genes Window—Correlations (Continued)

Spearman Confidence	Compute a value *r* of the spearman correlation as described above. *Result* =1-(probability you would get a value of *r* or higher by chance.)
Two-sided Spearman Confidence	Compute a value *r* of the spearman correlation as described above. *Result* =1-(probability you would get a value of |*r*| or higher, or -|r| or lower, by chance.)

Technical Details for Similarity Measures

Many advanced analysis techniques are based on measures of gene similarity. Similarity or "nearness" between genes is usually based on the correlation between the expression profiles of the two genes.

GeneSpring offers nine choices of similarity measures. Each can be selected from a pull-down list appearing in the Clustering and Filtering windows. See Chapter 15, "Clustering and Characterizing Data" and Chapter 9, "Filtering Data", respectively.

Each measure takes two expression patterns and produces a number representing how similar the two genes are. Most of the measures of similarity are correlation measures, and their value varies from -1 (exactly opposite) to 1 (the same). For a measure of distance, the result varies from 0 (the same) to infinity (different). For confidences, the result varies from 0 (no confidence) to 1 (perfect confidence). Both distance and confidence are actually measures of dissimilarity (small means close and large means far away). These are each transformed to measures of similarity by GeneSpring in ways detailed below.

If one expression value for a particular sample for either gene is missing, that sample is not considered in the calculation.

The notation used to describe the formulas:

- ***Result*:** the result of the calculation for genes **A** and **B**.
- ***n*:** the number of samples being correlated over.
- **a**: the vector $(a_1, a_2, a_3 \ldots a_n)$ of expression values for gene **A**.
- **b**: the vector $(b_1, b_2, b_3 \ldots b_n)$ of expression values for gene **B**.

Normal mathematical notation for vectors are used. In particular:

- **a.b** = $a_1b_1+a_2b_2+\ldots+a_nb_n$
- |**a**| = square root(**a.a**)

Standard Correlation

Standard correlation measures the angular separation of expression vectors for Genes **A** and **B** around zero. As almost all normalized values for genes are positive, you find mostly positive correlations between genes when you use the Standard correlation. This metric is designed to answers the question “do the peaks match up?” or to put it another way, “are the two genes expressed in the same samples?” Since these questions are the most frequent questions a biologist is trying to get answered, GeneSpring calls it “Standard correlation”. It is important to note, what mathematicians and statisticians refer to as “correlation” usually refers to the Pearson correlation. The “Standard correlation” would be called “Pearson correlation around zero” by mathematicians and statisticians.

To compute a Standard correlation:

Standard correlation = **a.b**/*(*|**a**||**b**|*)*

Figure 14-4 shows the summation notation.

$$\frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)}}$$

Figure 14-4 Summation Notation for the Standard Correlation

Pearson Correlation

The Pearson correlation is very similar to the Standard correlation, except that it measures the angle of expression vectors for genes **A** and **B** around the mean of the expression vectors (for example, the mean of the expression values constituting the profiles for Gene **A** and Gene **B**). Generally the mean of the expression vectors is positive since expression values are based on concentrations of mRNA. Using the Pearson correlation you get more negative correlations than from the Standard correlation (for example, you find more genes that behave opposite to each other, because of where you put the baseline—at zero almost all gene values are above it, at 1 there are a fair amount that read below the baseline).

For data normalized to an overall level of 1 (as with all normalizations that GeneSpring performs) the Pearson correlation gives you almost the same correlations as the Standard correlation when they are both performed on the logarithms of the genes’ expression values.

To compute a Pearson Correlation:

1. Calculate the mean of all elements in vector **a**.
2. Subtract that value from each element in **a**.
3. Do the same for **b**.

Pearson Correlation = **A.B**/*(*|**A**||**B**|*)*

Figure 14-5 shows the summation notation.

$$\frac{\sum_{i=1}^{n}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\left(\sum_{i=1}^{n}(A_i-\bar{A})^2\right)\left(\sum_{i=1}^{n}(B_i-\bar{B})^2\right)}}$$

Figure 14-5 Summation Notation for the Pearson Correlation

Spearman Correlation

The Spearman correlation is a non-parametric correlation. It is similar to the Pearson correlation, except it replaces the data for Gene **A** and **B** with the ranks of the data (for example the lowest measurement for a gene becomes 1, the second lowest 2, and so forth). Spearman correlation calculates the correlation of the ranks for Genes **A** and **B**'s expression data around the mean of the ranks, using the same formula as Pearson correlation.

In the Spearman correlation, the rank order of the data is import, not the level. Therefore, extreme variations in expression values have less control over the correlation. If there are ties in the data, then all of the tied values are assigned the average of the ranks. For example, if the 5th, 6th, and 7th lowest values are tied, all three data points are assigned a rank of 6.

To compute a Spearman correlation:

1. Order all the elements of vector **a**.
2. Use this order to assign a rank to each element of **a**.
3. Make a new vector **a'** where the i^{th} element in **a'** is the rank of a_i in **a**.
4. Now make a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation.
5. Similarly, make a vector **B** from **b**.

Spearman correlation = **A.B**/*(*|**A**||**B**|*)*

Spearman Confidence

Spearman confidence is a measure of similarity, not a correlation. Spearman confidence is one minus the p-value for the statistical test that the Spearman correlation is zero versus the alternative that is larger than zero.

A high Spearman confidence value occurs when you have a high Spearman correlation and a low p-value. Thus indicating that there is a low probability of finding a correlation this high. This measure is very to looking for high Spearman correlation values, but it takes into account the number of sub-experiments in your experiment set.

To compute a Spearman confidence:

If *r* is the value of the Spearman correlation as described in "Spearman Correlation" on page 14-17, then:

Spearman confidence =1-(probability you would get a value of *r* or higher by chance.)

Two-Sided Spearman Confidence

The two-sided Spearman confidence is a measure of similarity, but it is not a correlation. It is very similar to the Spearman confidence discussed in "Spearman Confidence" on page 14-18, except it is based on the two-sided test of whether the Spearman correlation is either significantly greater than zero or significantly lower than zero.

A high two-sided Spearman confidence value occurs when the absolute value of the Spearman correlation is large and has a small p-value. Thus indicating that there is a low probability of finding a correlation with an absolute value that is this large.

This similarity measure is really good for answering the question "What genes behave similarly to a specific gene, and at the same time, what genes behave opposite to a specific gene?". It should probably not be used for the advanced clustering algorithms (such as k-means and hierarchical clustering) because the genes with high two-sided confidence values are really a mixture of similar and dissimilar genes.

To compute a Two-sided Spearman confidence:

If *r* is the value of the Spearman correlation as described above, then:

Two-sided Spearman confidence =1-(probability you would get a Spearman correlation of |*r*| or higher, or -|r| or lower, by chance.)

Distance

Distance is not a correlation at all, but a measurement of dissimilarity. Distance is based on the measurement of euclidean distance between the expression profile for gene **A** (defined by its expression values for each point in n-dimensional space, where **n** is the number of experimental points (conditions) with data in your experiment) and the expression profile for gene **B**. This is more formally known as the euclidean metric. To standardize this difference GeneSpring divides by the square root of the number of conditions.

To compute a Euclidean distance:

Distance = |**a-b**| /square root of n
Since distance is a measure of dissimilarity, the distance (d) is converted when needed to a similarity measure 1/(1+d).

Figure 14-6 shows the summation notation.

$$D = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Figure 14-6 Summation Notation for Euclidean distance

The next three correlations should only be used to look at special cases. They are all modified versions of the Standard correlation. Using these three metrics only makes sense when your data are in a sequence, such as “before” and “after”, a time series, or a drug series. The sequence does not have to be continuous, but it must have an order. If your experiment is set up with an experimental point taken at each of “before”, “after”, and “control” the following correlations will not make sense applied to your data.

Smooth Correlation

To compute a Smooth correlation:

Make a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph window. Do the same to make a vector **B** from **b**.

Smooth correlation = **A.B**/*(*|**A**||**B**|*)*

Similarity between gene A and B

$$= \frac{((7 \cdot 3) + (2.5 \cdot 3.5) + (1.5 \cdot 5)(2.5 \cdot 8))}{((\sqrt{7^2 + 2.5^2 + 1.5^2 + 2.5^2})(\sqrt{3^2 + 3.5^2 + 5^2 + 8^2}))}$$

Experiment	1	2	3	4	5
Gene A	10	4	1	2	3
Gene B	2	4	3	7	9
Gene C	2	8	6	7	8
Between Experiments		1 and 2	2 and 3	3 and 4	4 and 5
Gene A		7	2.5	1.5	2.5
Gene B		3	3.5	5	8
Gene C		5	7	6.6	7.5

Change Correlation

The Change correlation looks for the opposite of what the Smooth correlation looks for. The change correlation only looks at the change in expression level of adjacent points. However, it is also very similar to the Standard correlation, in that it measures the angular separation of expression vectors for genes A and B around zero (for example, in comparison to zero). However, instead of using the expression values in each experimental point to create the expression vector for gene A, it is based on an arc tangent transformation of the ratio between adjacent pairs of experimental points. It uses these to create the expression vector.

This correlation looks for instances where gene A and gene B are changing at the same time. Using the arctangent makes a measure of change that is less sensitive to outliers than using the ratio directly.

To compute a Change correlation:

1. Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**.
2. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i)-\pi/4$.
3. Do the same to make a vector **B** from **b**.

Change correlation = $\mathbf{A}.\mathbf{B}/(|\mathbf{A}||\mathbf{B}|)$

Upregulated Correlation

The Upregulated correlation is very similar to the Change correlation, except that it only considers positive changes. All negative values for the arc tangent transform of the ratio are set to zero. This emphasizes only periods when new RNA is being synthesized.

To compute an Upregulated correlation:

1. Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**.
2. Do this for each pair of elements that would be connected by a line in the graph window. The value created between two values a_i and a_{i+1} is $\max(\text{atan}(a_{i+1}/a_i)-\pi/4,0)$.
3. Do the same to make a vector **B** from **b**.

Upregulated correlation = $\mathbf{A}.\mathbf{B}/(|\mathbf{A}||\mathbf{B}|)$

Number of Samples Required to do Analyses

For a gene to be included in an analysis listed below, it must have been assigned an expression value in at least half of the total number of conditions in the experiment. Each gene must also have the minimum number of measurements listed in Table 14-3.

Table 14-3 Required Number of Samples for Correlation Analyses

	k-means	SOM	Trees	Find Similar
Standard Correlation	2	N/A	2	2
Distance	1	2	1	1
Smooth Correlation	2	N/A	2	2
Change Correlation	3	N/A	3	3
Unregulated Correlation	3	N/A	3	3
Pearson Correlation	3	N/A	3	3
Spearman Correlation	3	N/A	3	3
Spearman Confidence	5	N/A	5	5
Two-Sided Spearman Confidence	5	N/A	5	5

For details on each of these clustering techniques, see Chapter 15, “Clustering and Characterizing Data”.

Performing Complex Correlations

To perform complex correlations:

1. Do one of the following:
   - Click on a gene in the browser window and select **Edit > Inspect Selected Gene**
   - Double-click a gene (this may be easier when you zoom in).
   - Select **Edit > Find Gene** and enter the name of the gene.
   - Press **Ctrl+I** (when one or more genes are selected).

   The Gene Inspector window opens.

   Go to “Using the Gene Inspector” on page 7-16 for more information.
2. Click **Complex Correlations**.

   The Find Similar Genes window opens.
3. In the **Navigator**, select the gene list you want, and click **Choose Gene List**.
4. Select a correlation from the **Similarity Measure** menu. The options are:
   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation

- Pearson Correlation
- Spearman Correlation
- Spearman Confidence
- Two-sided Spearman Confidence
- Distance

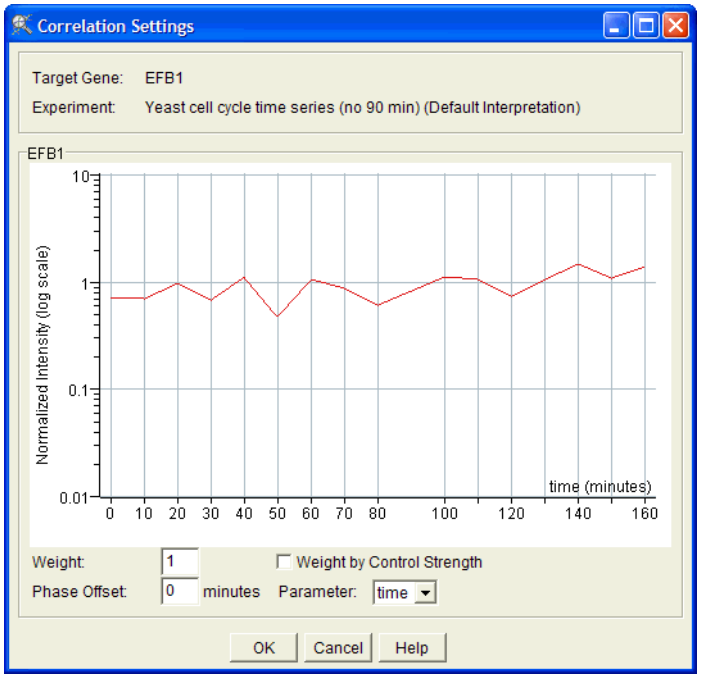
You can specify minimum and maximum settings for the similarity measure by moving the sliders or entering values in the appropriate fields.

5. Select an experiment or condition in the Navigator and click **Add**.

   By default, the currently selected experiment will be used to find genes with a similar expression profile. If you would like to add more experiments in which to find similar expression profiles, you can add more.

   For each added experiment, the gene expression profile for the selected gene in each experiment will be compared with the gene expression profiles for each other gene only within the same experiment. This will allow you to find genes that have a similar expression profiles under varying conditions, and will provide you with strong and independent evidence for co-regulation or co-expression.

   The New Correlation window opens.



6. (Optional) Enter new values for Phase Offset and Weight. You can select a parameter from the menu in the phase offset section.
7. Click the **OK** button.

   The Find Similar Genes window opens.

8. Do one of the following:
   - To add additional experiments or conditions, go to "Correlations Table" on page 14-13.
   - To remove an experiment or condition, click on the name of the experiment or condition in the white center box and click the **Remove** button.
   - To change the settings, click the experiment's name to select its row and click the **Edit** button.
9. Specify whether to run the search locally or on a Signet server.
10. When you are done, click the **OK** button.

    The New Gene List window opens.
11. Name your gene list and click the **Save** button.

    The list appears in the Gene Lists folder of the main Navigator.

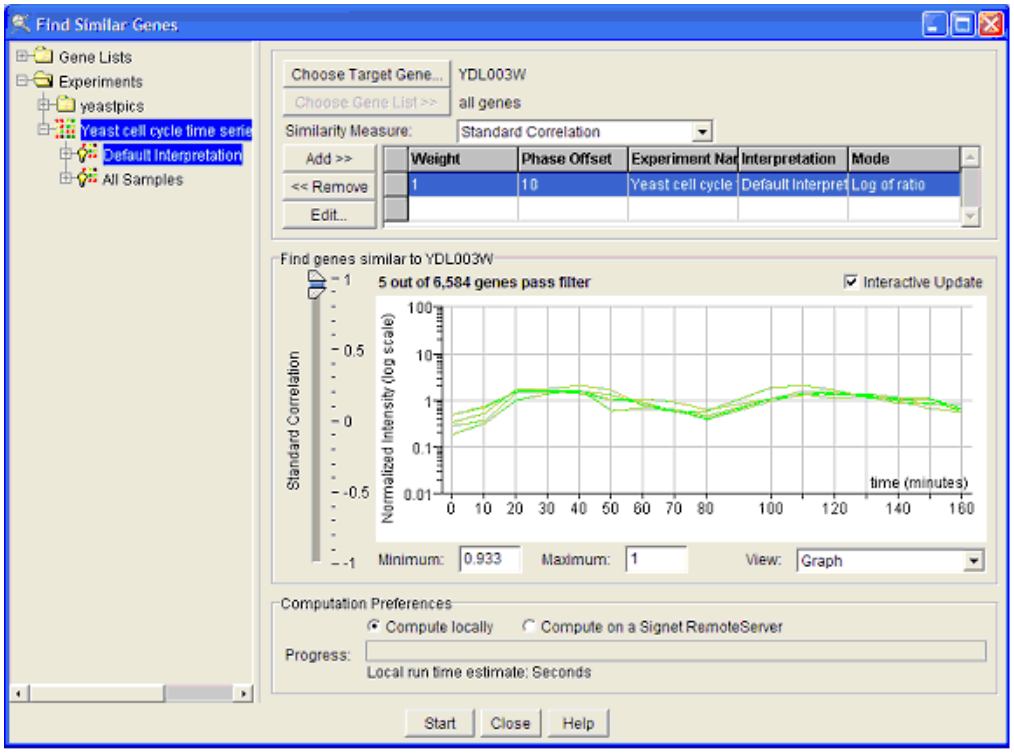
Finding Similar Genes Based on Correlations

Using the Find Target Gene window, you can search for specific genes, called "target" genes using a similarity measure, and then build a gene list from the target genes.

To make a gene list from target genes:

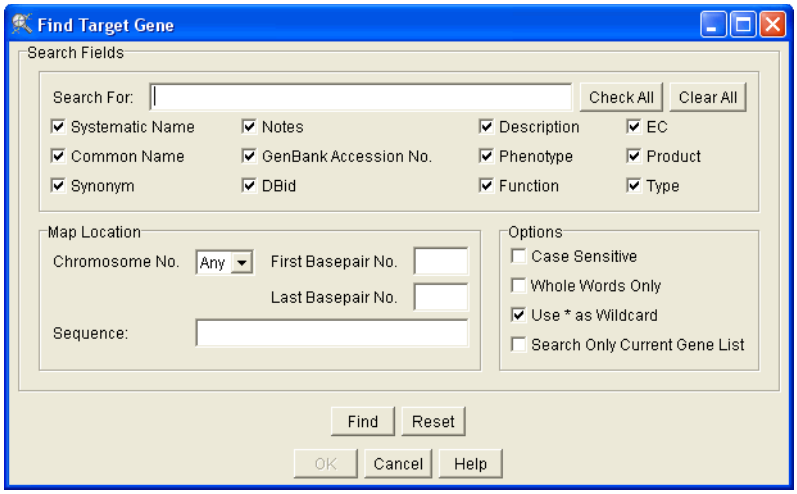
1. Select **Tools > Find Similar Genes**.

   The Find Similar Genes window opens.



2. Click the **Choose Target Gene** button.

   The Find Target Gene window opens.



3. In the **Annotation** check boxes, do one of the following:
   - To search individual annotations, select the check box you want.
   - To search all annotations, click the **Check All** button.
   - To clear all selected annotations, click **the Clear All** button.
4. In the **Search For** field, enter the search terms you want:
   - Create a Boolean search by entering multiple terms separated by the words “and”, “or”, or “not” in this field.
     - **AND**—Matches any gene containing both of these terms, even if they do not appear in the same field.
     - **OR**—Matches any gene containing either of these terms, even if they do not appear in the same field.
     - **NOT**—Matches any gene containing the first term, but not containing the second.

     White space in the Search For field behaves as an AND between words or phrases.
   - Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified.
   - You can restrict your search to specific regions within the genome using the **Map Location** fields. You can specify the chromosome number or search for a genome sequence.
   - To perform an exact string match, use double quotes in your search term. For example, “expression profile”.
   - For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given

genome appear. Any gene that falls even partially within the specified region is identified by the search.

- You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.

5. To restrict your search, you can select any of the following options:

   - **Case Sensitive**—Searches only for words using the specific capitalization you entered.
   - **Whole Words Only**—Searches only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
   - **Use * as a Wildcard**—Searches for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, for example, “ex*.
   - **Whole Words Only** and **Use * as a Wildcard**—Performs a search using Whole Words Only combined with Use * as a Wildcard, with the following provisions:
     - Any words that do not have an * next to them will have the Whole Words Only search applied.
     - Words with an * in them or next to them will not have the Whole Words Only Search applied.
     - If there is no * in the search field, then the Use * as Wildcard box functions as if it is not selected.
   - **Search Only Current Gene List**—Searches the current gene list only.

6. Click the **Find** button.

   A list of the genes that match the search criteria is displayed in a table at the bottom of the window. The field that matched the search criteria is colored red.

7. To learn more about one or more genes, do one of the following:

   - Select the gene of interest in the appropriate row.
   - To select more than one gene, hold down the **Ctrl** key while clicking multiple rows.

8. Once you have highlighted the gene(s) of interest, do the following:

   - To select the genes in the Genome Browser, click the **Select** button.
   - To select all of the genes identified by the search, click **Select All** button.
   - To select and zoom-in on a gene in the Genome Browser, click the **Zoom & Select** button.
   - To display the Gene Inspector window for the selected gene, click the **Inspect** button.

- To make a gene list from all of the genes identified by the search, click the **Make Gene List** button.

9. Click the **Find** button.
10. When you are done searching, click the **OK** button.

    Results appear at the bottom of the window.
11. When you are done, click the **Close** button.

15

Clustering and Characterizing Data

This chapter explains how to find similar expression patterns using different clustering and data characterization techniques. It covers the following topics:

- Clustering
- Performing Clustering Analysis
- Performing Principal Components Analysis
- Performing Class Prediction Analysis
- Finding Significant Parameters

Clustering

GeneSpring provides sophisticated clustering methods to:

- Uncover patterns of gene expression data and the relationships between these patterns
- Reduce the complexity of your data
- Discover genes that are primarily responsible for the variation

You can use one or a combination of clustering options to characterize their data: gene trees (hierarchical clustering), condition trees, self-organizing maps, k-means, Principal Components Analysis (PCA) and QT clustering.

Principal Components Analysis (PCA), allows you to reduce the complexity of your data by discovering a number of principal components that define most of the data variability. QT clustering is an unsupervised technique that allows you to specify both the minimum size and maximum correlation coefficient of each cluster.

Performing Clustering Analysis

GeneSpring's clustering algorithms are designed to divide genes or conditions into groups that have similar expression patterns. GeneSpring supports a variety of clustering methods, each designed to solve a distinct type of problem. These are useful tools to identify genes that are potentially co-regulated as well as to reveal coordinated responses to experimental treatments.

Clustering Window

The Clustering window (Figure 15-1) lets you perform the following clustering analyses:

- K-means
- Gene Tree
- Condition Tree
- Self-Organizing Map
- QT Clustering

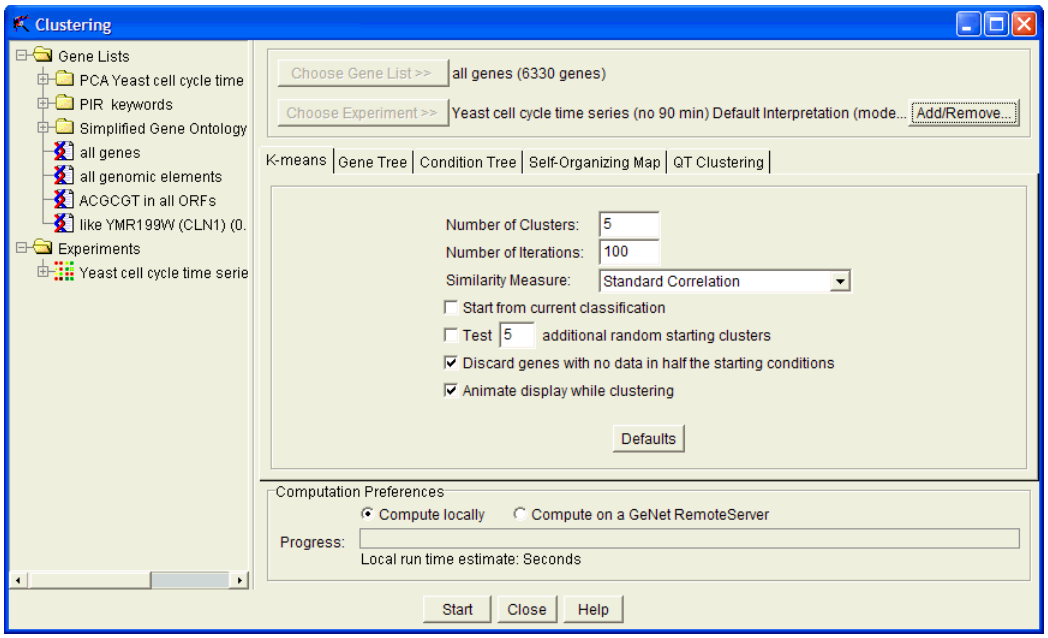


Figure 15-1 Clustering Window

K-Means Clustering

K-means clustering divides genes into groups based on their expression patterns. The goal is to produce groups of genes with a high degree of similarity within each group and a low degree of similarity between groups. Unlike self-organizing maps, k-means clustering is not designed to show the relationship between clusters. Instead, k-means clusters are constructed so that the average behavior in each group is distinct from any of the other groups. For example, in a time series experiment you could use k-means clustering to identify unique classes of genes that are upregulated or downregulated in a time dependent manner.

GeneSpring's k-means clustering algorithm divides genes into a user-defined number (k) of equal-sized groups, based on the order in the selected gene list. It then creates centroids (in expression space) at the average location of each group of genes. With each iteration, genes are reassigned to the group with the closest centroid. After all of the genes have been reassigned, the location of the centroids is recalculated and the process is repeated until the maximum number of iterations has been reached.

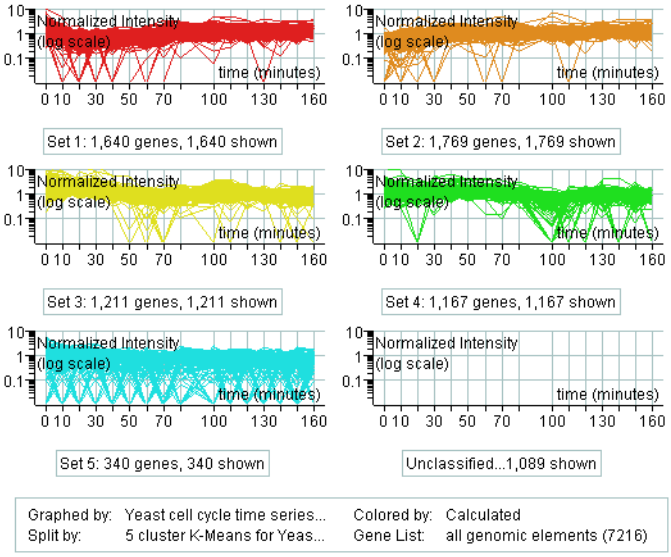

Normalized Intensity (log scale)
time (minutes)
Set 1: 1,640 genes, 1,640 shown
Set 2: 1,769 genes, 1,769 shown
Set 3: 1,211 genes, 1,211 shown
Set 4: 1,167 genes, 1,167 shown
Set 5: 340 genes, 340 shown
Unclassified...1,089 shown
Graphed by: Yeast cell cycle time series...
Split by: 5 cluster K-Means for Yeas...
Colored by: Calculated
Gene List: all genomic elements (7216)

Figure 15-2 K-means Cluster Display in a Split Window

Classifications

The Classifications folder in the Navigator contains genes that have been grouped or *classified* into groups as defined by K-means or SOM clustering.

Gene Tree

The classification of organisms into phylogenetic trees is a central concept to biology. Organisms sharing properties tend to be clustered together, and the location of a branch containing both organisms can be considered a measure of how similar the organisms are. You can classify genes in a similar manner—clustering those whose expression patterns are similar into nearby places in a tree. Such mock-phylogenetic trees are often referred to as gene trees.

GeneSpring can both create and display such trees. GeneSpring can also create trees of experiments, displaying the genes along one axis and the samples along the other axis. This is useful for many applications. For example, you can determine if any environmental stressors cause similar effects on the expression levels as mutant organisms do.

Any gene trees created in GeneSpring are kept in the Gene Trees folder. Gene trees are dendrograms used as a method of showing relationships among the expression levels of genes over a series of conditions.

If you have already created or downloaded trees, open the Gene Trees folder in the Navigator and select any tree for viewing.

Condition Tree

Complex trees can be made from multiple conditions or by tightly defining the types of data to use. Select a gene list in the Navigator to reduce the number of genes to be made into a tree.

Condition trees are like gene trees, except that instead of showing the relationships between genes, they show the relationships among the expression levels of samples. Condition trees are kept in the Condition Trees folder.

Self-Organizing Maps

The self-organizing map (SOM) is a clustering technique similar to k-means clustering. However, SOMs illustrate the relationship between groups by arranging them in a two-dimensional map in addition to dividing genes into groups based on expression patterns. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another. SOMs were invented by Tuevo Kohonen (1991, 2000) and are used to analyze many kinds of data. Applications to gene expression analysis were described by Tamayo, et. al. (1999).

GeneSpring's self-organizing map algorithm begins by creating a two-dimensional grid of nodes in the space of gene expression. In each iteration, one gene is selected and all of the nodes within a user-defined "neighborhood" are moved closer to it. This process is repeated with each gene in the selected gene list until the maximum number of iterations has been reached.

With each iteration, the "neighborhood radius" is incrementally reduced and nodes are moved by smaller and smaller amounts to produce convergence. In this way, the grid of nodes is stretched and wrapped to best represent the variability of the data, while still maintaining similarity between adjacent nodes. After the iteration is complete, genes are assigned to the nearest node, and a display grid of gene expression graphs is generated, corresponding to the initial grid of nodes.

As the iteration proceeds, the neighborhood radius decreases smoothly, so that points move more independently later in the process. The neighborhood radius is expressed in terms of Euclidean distance in grid units relative to the abstract grid of the expression patterns. (This is different from the distance between nodes in gene expression space.) For instance, point 1,2 is one unit away from 1,3.

If you make the neighborhood radius very small (less than 1) each point always moves independently, and adjacent clusters are not related. If you specify a very large neighborhood radius, initially all the nodes move toward every data point, and the grid behaves as if it is very "stiff", with more similarity between node results, but less flexibility to explore the variations in the data.

QT Clustering

QT clustering looks for clusters of genes such that each gene in the cluster is within a specified distance (based on a user-defined distance metric) of every other gene in the cluster. In GeneSpring, the cutoff is specified based on a correlation function, so the cutoff is the minimum allowed value: In QT clustering, the “diameter” of a cluster refers to the largest distance between any two genes in the same cluster.

QT clustering builds a cluster by starting with a single gene. The diameter at that point is 0. It then adds the gene that is closest to the starting gene. The diameter of the cluster is now equal to the distance between the two genes. It continues adding genes one at a time, always choosing the gene that will result in the smallest cluster diameter. Eventually it reaches a point where no genes can be added without the diameter growing beyond the allowed cutoff. The cluster is then complete.

The cluster obtained depends on which gene is chosen to start from. Therefore, it independently builds clusters starting from each gene in the user-selected gene list. The cluster with the most genes is kept, and is part of the final classification. All others are discarded. A new cluster is built from every gene in the reduced gene list, where the largest one is kept. This process is repeated until the number of genes in the largest cluster is smaller than a user-defined cutoff.

Performing K-means Clustering

To perform k-means clustering:

1. Select **Tools > Clustering > K Means**.

   The Clustering window opens.

2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. Click **Add/Remove** to add or remove experiments from the list to be analyzed.

   For more information on adding and removing experiments, see “Adding or Removing Experiments” on page 15-19.

5. Enter settings for the clustering operation. The following options are available:
   - **Number of Clusters**—The number of clusters to make.
   - **Number of Iterations**—The maximum number of times that each centroid is recalculated after genes are reassigned to groups with the most similar centroids.
6. In the **Similarity Measures** list, select the type of correlation you want. The options are:
   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation
   - Spearman Confidence
   - Two-sided Spearman Confidence
   - Distance

   For more information on measures of similarity, see "Similarity Measures" on page 15-20.
   - **Start from Current Classification**—Group genes using the selected classification as a starting point. Note that this option is available only if you have selected a classification. This option disables the Number of Clusters checkbox, since it automatically uses the number of classes in the current classification.
   - **Test [*x*] Additional Random Starting Clusters—** Enter a number to make clustering as tight as possible by performing clustering several times, each time starting from a different random grouping of genes, and choosing the best result. The default value is 5.
   - **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment. These genes will be placed in the Unclassified cluster.
   - **Animate display while clustering**—Show changes in classification assignments in real time. This action may slow your analysis slightly.
7. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.
8. Click the **Start** button.

   If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the window.

Saving K-means Clustering Results

When the k-means operation is complete, the Save New Classification window (Figure 15-3) opens.

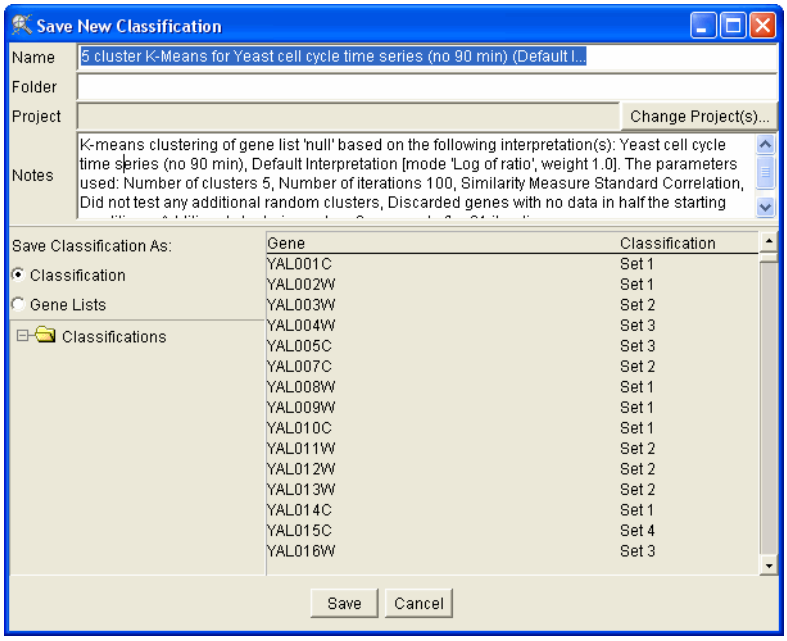

Save New Classification
Name
5 cluster K-Means for Yeast cell cycle time series (no 90 min) (Default I...
Folder
Project
Change Project(s)...
Notes
K-means clustering of gene list 'null' based on the following interpretation(s): Yeast cell cycle time series (no 90 min), Default Interpretation [mode 'Log of ratio', weight 1.0]. The parameters used: Number of clusters 5, Number of iterations 100, Similarity Measure Standard Correlation, Did not test any additional random clusters, Discarded genes with no data in half the starting
Save Classification As:
Classification
Gene Lists
Classifications
Gene
Classification
YAL001C Set 1
YAL002W Set 1
YAL003W Set 2
YAL004W Set 3
YAL005C Set 3
YAL007C Set 2
YAL008W Set 1
YAL009W Set 1
YAL010C Set 1
YAL011W Set 2
YAL012W Set 2
YAL013W Set 2
YAL014C Set 1
YAL015C Set 4
YAL016W Set 3
Save
Cancel

Figure 15-3 Save New Classification Window

To save k-means clustering results:

1. Enter a name in the **Name** field at the top of the window.

   Names cannot exceed 80 characters.

2. Do one of the following:

   - To save the results as a classification, select the **Classification** option.
   - To save the results as a group of gene lists, select the **Gene Lists** option.

3. In the **Navigator**, select a folder in which to save the new classification or gene lists.
4. To create a new folder, navigate to the parent folder you want and enter a new folder name in the **Folder** field.
5. Enter any additional information in the **Notes** field you want.
6. Click the **Save** button.

Viewing K-means Cluster Results

If you use k-means clustering to produce a classification, you can view details about the classification in the Classification Inspector. For information about the Classification Inspector, see "Using the Classification Inspector" on page 7-37.

The easiest way to view a classification is to simply select the classification you want.

To view k-means clustering results:

- Do one of the following:
  - Left-click the classification you want.
  - Right-click the classification you want and select **Split Window**.

Performing Gene Tree Clustering

To perform gene tree clustering:

1. Select **Tools** > **Clustering** >**Gene Tree**.

   The Clustering window opens.

2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. Click **Add/Remove** to add or remove experiments from the list to be analyzed.

   For more information on adding and removing experiments, see "Adding or Removing Experiments" on page 15-19.

5. Enter settings for the clustering operation. The following options are available:

6. In the **Similarity Measures** list, select the type of correlation you want. The options are:

   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation

- Spearman Confidence
- Two-sided Spearman Confidence
- Distance

For more information on measures of similarity, see “Similarity Measures” on page 15-20.

- **Do automatic annotation**—Specifies whether to annotate the nodes of the tree with the names of the gene lists that have similar members. The automatic annotation feature will look for gene lists that have significant overlap with the genes in a particular node of the tree, which will be an indication that the genes in those nodes are significantly enriched with genes from those gene list.

  Using this option can add considerable time to the tree-building process, but is usually worthwhile.This feature becomes even more valuable once you have created a simplified ontology for the genome, as the ontological classifications can be used to label tree branches.

  For details on creating a simplified ontology, see “Working with Homology Tables” on page 12-23.

- **Only annotate with standard lists**—Specifies whether the annotations on the nodes are done with all gene lists or only the gene lists marked as standard. (This is set in the Gene List Inspector. See “Using the Gene List Inspector” on page 7-33 for more information.)
- **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment.These genes will be placed in the Unclassified cluster.
- **Merge similar branches**—Merge branches with similar results. For information on the Separation Ratio and Minimum Distance settings, see “Advanced Tree Options” on page 15-13.

7. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

8. Click the **Start** button.

   If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the window.

Saving Gene Tree Clustering Results

When the gene tree clustering operation has completed, the Name New Gene Tree window (Figure 15-4) opens.

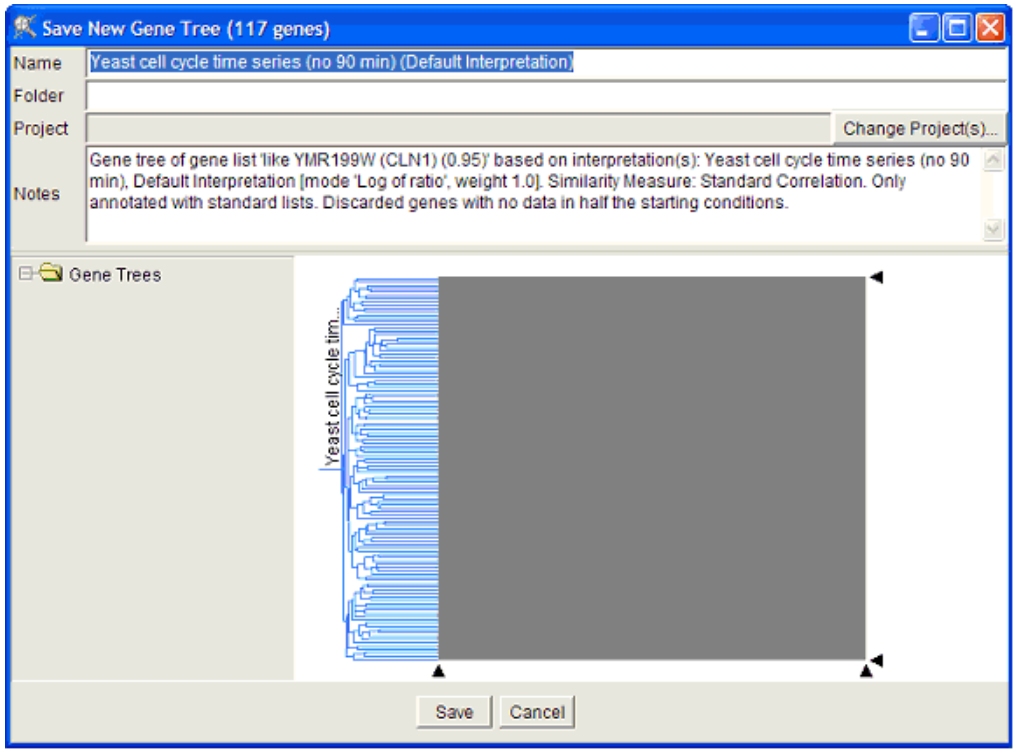

Save New Gene Tree (117 genes)
Name
Yeast cell cycle time series (no 90 min) (Default Interpretation)
Folder
Project
Change Project(s)...
Notes
Gene tree of gene list 'like YMR199W (CLN1) (0.95)' based on interpretation(s): Yeast cell cycle time series (no 90 min), Default Interpretation [mode 'Log of ratio', weight 1.0]. Similarity Measure: Standard Correlation. Only annotated with standard lists. Discarded genes with no data in half the starting conditions.
Gene Trees
Yeast cell cycle tim...
Save
Cancel

Figure 15-4 Name New Gene Tree Window

To save gene tree clustering results:

1. Enter a name in the **Name** field at the top of the window.

   Names cannot exceed 80 characters.

2. In the **Navigator**, select a folder in which to save the new gene tree.
3. To create a new folder, navigate to the parent folder you want and enter a new folder name in the **Folder** field.
4. Enter any additional information in the **Notes** field you want.
5. Click the **Save** button.

Performing Condition Tree Clustering

To perform condition tree clustering:

1. Select **Tools** > **Clustering** > **Condition Tree**.

   The Clustering window opens.

2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Gene List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. Click **Add/Remove** to add or remove experiments from the list to be analyzed.

   For more information on adding and removing experiments, see “Adding or Removing Experiments” on page 15-19.

5. In the **Similarity Measures** list, select the type of correlation you want. The options are:

   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation
   - Spearman Confidence
   - Two-sided Spearman Confidence
   - Distance

     For more information on measures of similarity, see “Similarity Measures” on page 15-20.

   - **Merge similar branches**—Merge branches with similar results. For information ion the Separation Ratio and Minimum Distance settings, see “Advanced Tree Options” on page 15-13.

6. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

7. Click the **Start** button.

   If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the window.

Saving Condition Tree Results

When the operation is complete, the Save New Condition Tree window (Figure 15-5) opens.

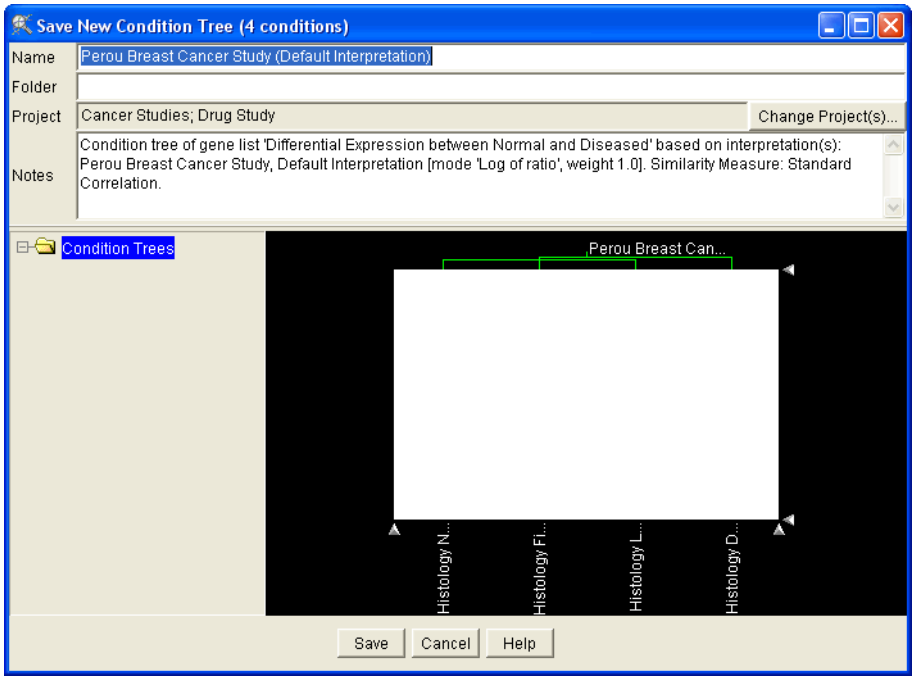


Figure 15-5 Save New Condition Tree Window

To save condition tree clustering results:

1. Enter a name in the **Name** field at the top of the window.

   Names cannot exceed 80 characters.

2. In the **Navigator**, select a folder in which to save the new condition tree.

3. To create a new folder, navigate to the parent folder you want and enter a new folder name in the **Folder** field.

4. Enter any additional information in the **Notes** field you want.

5. Click the **Save** button.

Advanced Tree Options

The separation ratio determines how large the correlation difference between groups of clustered genes must be for them to be considered discrete groups. This number should be between 0 and 1.It is not usually appropriate to change separation ratio or minimum distance.

Separation Ratio

The separation ratio determines how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

- Increasing separation increases the ‘branchiness’ of the tree.
- Default Separation ratio is 1.0. Separation ratio can range from 0.0 to 1.0.
- At a separation ratio of 0, all gene expression profiles can be regarded as identical.

To change the maximum correlation number, enter a new value in the Separation Ratio box.

Minimum Distance

The number specified in the Minimum distance box determines the minimum separation considered significant between genes. This reduces meaningless structure at the base of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A higher number tends to lump more genes into a group, making the groups less specific.

- Decreasing minimum distance increases the ‘branchiness’ of the tree.
- Default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely.

To change the default minimum distance, enter a new value in the Minimum distance box.

Viewing Condition Tree Clustering Results

To view condition tree clustering results:

1. In the **Navigator**, open the **Gene Tree** or **Condition Tree** folder you want.
2. Double-click the tree you want to view.

References for Hierarchical Clustering

Everitt, Brian S. *Cluster Analysis* (3rd Ed.) Arnold, London, 1993, pp 62-65.

Eisen, Michael B., et. al. “Cluster analysis and display of genome-wide expression patterns” *Proc. Natl. Acad. Sci. USA*, V95, pp 14863-14868, December 1998.

Performing Self-Organizing Map Clustering

To perform SOM clustering:

1. Select **Tools** > **Clustering** > **Self-Organizing Map**.

   The Clustering window opens.

2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. Click **Add/Remove** to add or remove experiments from the list to be analyzed.

   For more information on adding and removing experiments, see "Adding or Removing Experiments" on page 15-19.

5. Enter settings for the clustering operation. The following options are available:

   - **Rows**—The number of rows in your grid. The default setting is based on the number of genes and conditions in the selected experiment(s).
   - **Columns**—The number of columns in your grid. The default setting is based on the number of genes and conditions in the selected experiment(s).
   - **Number of Iterations**—How many times each gene is examined. For example, if there are 10,000 genes and 60,000 iterations are specified, each gene is examined six times.
   - **Neighborhood Radius**—How many nodes move toward a data point at the beginning of the iteration, and therefore how similar the profiles are for each node.
   - **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment.These genes will be placed in the Unclassified cluster.

     Note: A good way to estimate the optimum number of rows and columns is to try to predict how many distinct classes of genes are affected by the conditions in your experiment. With small data sets, the algorithm may generate a number of empty nodes. To avoid this, you might try using a smaller grid.

6. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

7. Click the **Start** button.

   If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the window.

Saving Self-Organizing Map Clustering Results

When the SOM operation is complete, the Save New Classification window (Figure 15-6) opens.

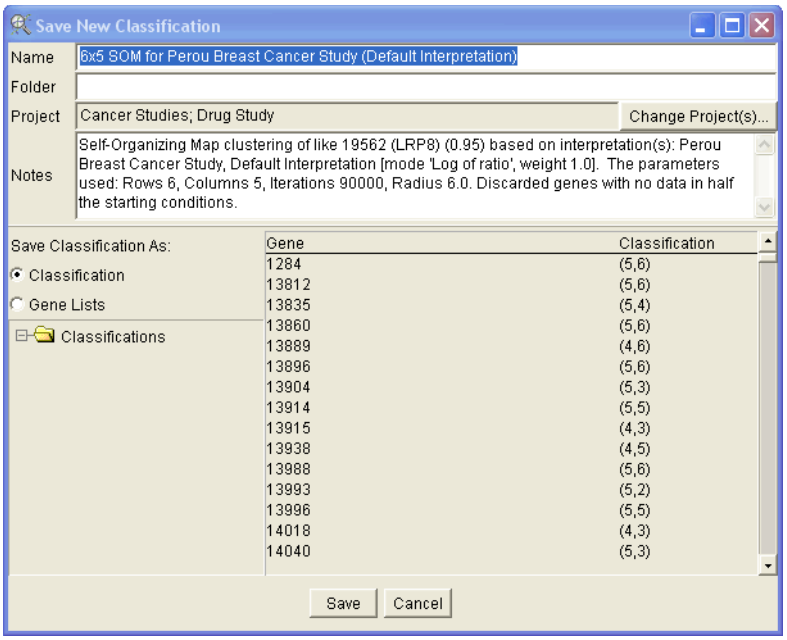


Figure 15-6 Save New Classification Name Window

To save SOM clustering results:

1. Enter a name in the **Name** field at the top of the window.

   Names cannot exceed 80 characters.

2. Do one of the following:

   - To save the results as a classification, select the **Classification** option.
   - To save the results as a group of gene lists, select the **Gene Lists** option.

3. In the **Navigator**, select a folder in which to save the new classification or gene lists.
4. To create a new folder, navigate to the parent folder you want and enter a new folder name in the **Folder** field.
5. Enter any additional information in the **Notes** field you want.
6. Click the **Save** button.

Viewing Self-Organizing Map Clustering Results

SOM results are most easily viewed using the Split Window feature. Each graph contains the genes associated with a SOM node (Figure 15-7).

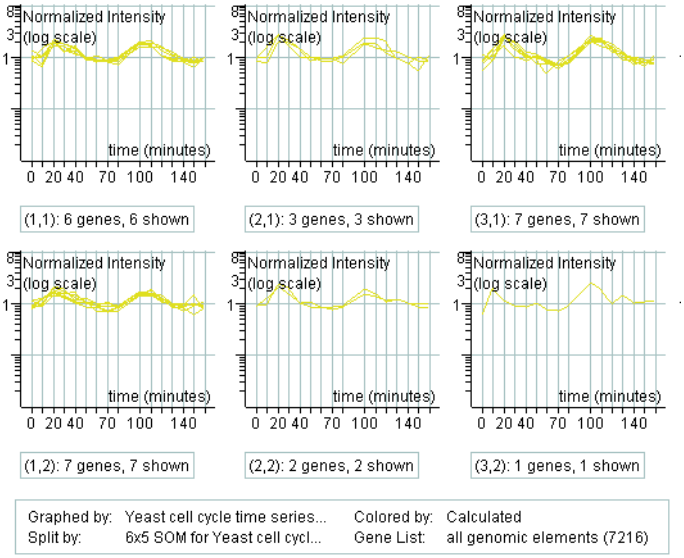

Normalized Intensity (log scale)
time (minutes)
0 20 40 70 100 140
(1,1): 6 genes, 6 shown
(2,1): 3 genes, 3 shown
(3,1): 7 genes, 7 shown
(1,2): 7 genes, 7 shown
(2,2): 2 genes, 2 shown
(3,2): 1 genes, 1 shown
Graphed by: Yeast cell cycle time series...
Split by: 6x5 SOM for Yeast cell cycl...
Colored by: Calculated
Gene List: all genomic elements (7216)

Figure 15-7 3x2 SOM of the "Yeast Cell Time Series (No 90 Min)" Experiment

If you have selected many panels, you may want to hide the horizontal and vertical labels for easier viewing.

To view SOM clustering results:

1. In the **Genome Browser**, right-click and select an option from the **Options** menu.
2. You can also increase your viewing space by selecting **View > Visible > Hide All**.
3. If you use a SOM to produce a classification, you can get details about the classification from the Classification Inspector.
4. For information about the Classification Inspector, see "Using the Classification Inspector" on page 7-37.
5. To recreate your SOM graph, click the SOM classification or folder of gene lists in the Navigator and select **Split Window > Both**.

SOM References

Kohonen, T. (1990). The Self-Organizing Map. *Proc*. IEEE 78(9):1464-1480.

Kohonen, T. (2000). *Self-Organizing Maps* (Third Edition). Springer Verlag. Berlin.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps; Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci.* USA 96:2907-2912.

Performing QT Clustering

1. Select **Tools > Clustering > QT Clustering**.

   The Clustering window opens.

2. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

3. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

4. Click **Add/Remove** to add or remove experiments from the list to be analyzed.

   For more information on adding and removing experiments, see "Adding or Removing Experiments" on page 15-19.

5. Enter settings for the clustering operation. The following options are available:

   - **Minimum Cluster Size**—The smallest allowable size for a cluster to be considered valid.
   - **Minimum Correlation**—The minimum correlation for any pair of genes in the same cluster.

6. In the **Similarity Measures** list, select the type of correlation you want. The options are:

   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation
   - Spearman Confidence
   - Two-sided Spearman Confidence
   - Distance

   For more information on measures of similarity, see "Similarity Measures" on page 15-20.

7. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

8. Click the **Start** button.

   If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the window.

Saving QT Clustering Results

When the operation is complete, the Save New Classification window (Figure 15-8) appears.

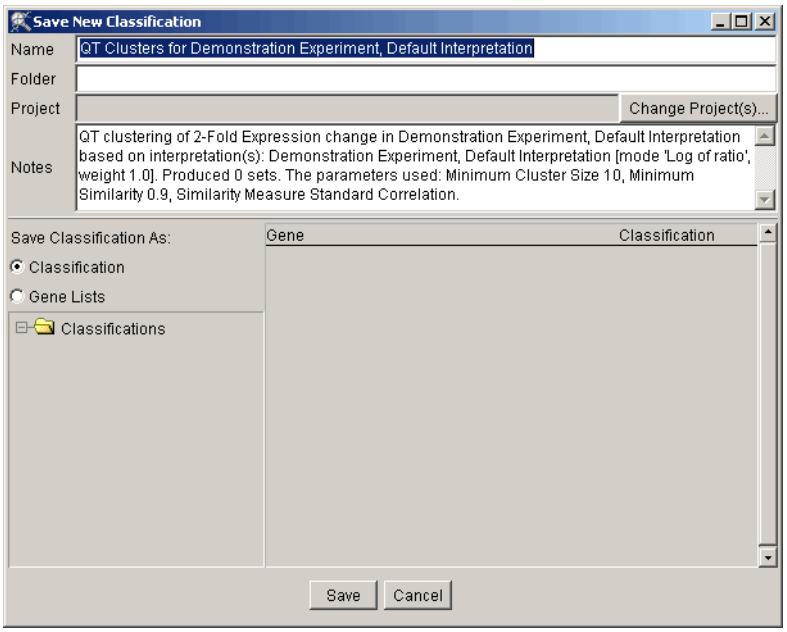

Save New Classification
Name
QT Clusters for Demonstration Experiment, Default Interpretation
Folder
Project
Change Project(s)...
Notes
QT clustering of 2-Fold Expression change in Demonstration Experiment, Default Interpretation based on interpretation(s): Demonstration Experiment, Default Interpretation [mode 'Log of ratio', weight 1.0]. Produced 0 sets. The parameters used: Minimum Cluster Size 10, Minimum Similarity 0.9, Similarity Measure Standard Correlation.
Save Classification As:
Classification
Gene Lists
Classifications
Gene
Classification
Save
Cancel

Figure 15-8 Save New Classification Window

To save QT clustering results:

1. Enter a name in the **Name** field at the top of the window.

   Names cannot exceed 80 characters.

2. Do one of the following:

   - To save the results as a classification, select the **Classification** option.
   - To save the results as a group of gene lists, select the **Gene Lists** option.

3. In the **Navigator**, select a folder in which to save the new classification or gene lists.
4. To create a new folder, navigate to the parent folder you want and enter a new folder name in the **Folder** field.
5. Enter any additional information in the **Notes** field you want.

6. Click the **Save** button.

Adding or Removing Experiments

Use the Experiments to Cluster window (Figure 15-9) to add or remove experiments from the list to be included in the clustering operation.

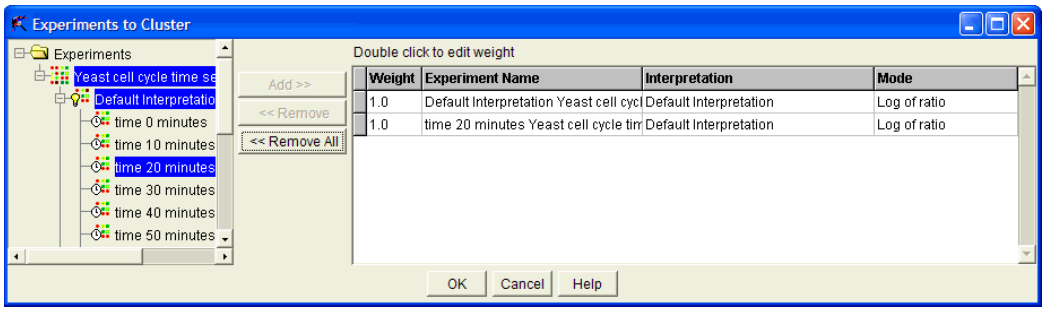


Figure 15-9 Experiments to Cluster Window

To add or remove an experiment:

1. To add an experiment, select it from the **Navigator** and then click the **Add** button.
2. To remove an experiment, select it in the list and then click the **Remove** button.
3. To remove all listed experiments, click the **Remove All** button.
4. To change an experiment's weight, double-click its number in the **Weight** column of the **Experiments to Cluster window** and enter a new value.

   Go to "Experiment Weight" on page 15-19 for more information.
5. When you are done adding or removing experiments, click **OK** to return to the main clustering window.

Experiment Weight

Correlations of multiple experiments are performed through a weighted correlation in which you specify the weight of each experiment. You can make one experiment more important than another. If all of the experiments or experiment sets are given the same weight, they are averaged equally.

The name of the experiment is noted directly after its relative weight. For example, you could give SampleExperiment1 a weight of 2, and Experiment2 a weight of 1. Therefore, in this example, the correlations found in the SampleExperiment1 are twice as influential in creating the tree as the correlations between the genes in the Experiment2 study.

The equation used to determine the overall correlation is:

$$\frac{X = (Aa + Bb + Cc + \ldots)}{(a + b + c + \ldots)}$$

The variables are:

- **A**—The correlation coefficient between the gene in question in experiment 1 and the gene named in the *Experiments to Use* box, also from experiment 1.
- **a** is the weight specified for experiment 1.
- **B**—The correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- **b**—The weight associated with experiment 2.
- **C**—The correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- **c**—The weight associated with experiment 3.

  and so on.

Experiments 1, 2, 3, etc., represent all of the experiments selected in the white Correlations box. If **X** is between the minimum and maximum correlations specified in the Clustering window, the gene in question passes the correlations.

Similarity Measures

Similarity measures are used in several clustering types. The equations used to determine the nine types of correlations are described in detail in "Correlation of Expression Change" on page 14-11.

The default correlation is the Standard Correlation, *Standard correlation* = **a.b**/(|**a**||**b**|).

The minimum distance and separation ratios equation is:

$$\frac{\sum_{i-1}^{n} A_i B_i}{\sqrt{\left(\sum_{i-1}^{n} A_i^2\right)\left(\sum_{i-1}^{n} B_i^2\right)}}$$

To make a tree, GeneSpring calculates the correlation for each gene with every other gene in the set. Then it takes the highest correlation and pairs those two genes, averaging their expression profiles. GeneSpring then compares this new composite gene with all of the other unpaired genes.

This is repeated until all of the genes have been paired. At this point the minimum distance and the separation ratio come in to play. Both of these affect the branching behavior of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A value smaller than .001 has very little effect, because most genes are not correlated more closely than that. A higher number tends to lump more genes into a group, making the groups less specific.

Performing Principal Components Analysis

Principal components analysis (PCA) is a decomposition technique that produces a set of expression patterns known as principal components. Linear combinations of these patterns can be assembled to represent the behavior of all of the genes in a given data set.

PCA is not a clustering technique. It is a tool to characterize the most abundant themes or building blocks that reoccur in many genes in your experiment. You can run PCA on genes or on conditions.

By default for PCA on Genes, PC scores are calculated by computing the standard correlation between each gene's expression profile vector and each principal component vector (eigenvector). For PCA on conditions, this means calculating the standard correlation between each condition vector and each principal component vector (eigenvector). Calculating scores this way has the advantage of scaling them to be between -1 and 1.

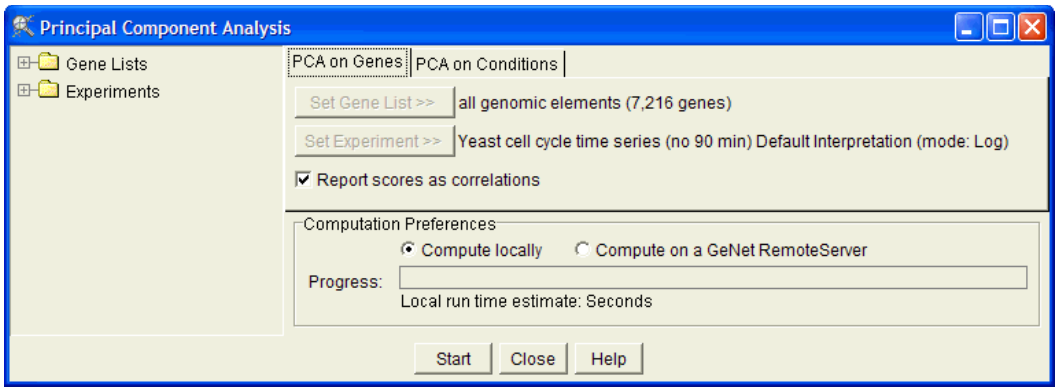
If you clear the **Report scores as correlations** box, the PC scores represent the coordinates of the genes or conditions in the system defined by the first few principal components. In other words, these scores are the values of the principal components for each gene or condition.

Running a PCA on Genes

To run a PCA on genes:

1. Select **Tools > Principal Components Analysis**.

   The Principal Components Analysis window opens.



2. Click the **PCA on Genes** tab.
3. In the **Navigator**, open the **Gene List** folder, select the gene list you want, and then click the **Set Gene List** button.
4. Select an experiment from the Navigator and click the **Set Experiment** button.
5. Select or clear the **Report scores as correlations** box to specify whether to report scores as correlations or as the values of the principal components for each gene. This box is selected by default.

6. Specify whether to run the computation locally or on a Signet server.
7. Click the **Start** button.

Viewing PCA on Genes Results

When the analysis is complete, the PCA Results window (Figure 15-10) opens. It displays each component as a line in graph mode. The significance of each component is represented by the color of its graph line, as defined by the Colorbar.

In addition, a new gene list folder appears in the GeneSpring Navigator with a name that includes the experiment that you used for PCA analysis (for example, “PCA yeast cell cycle”).

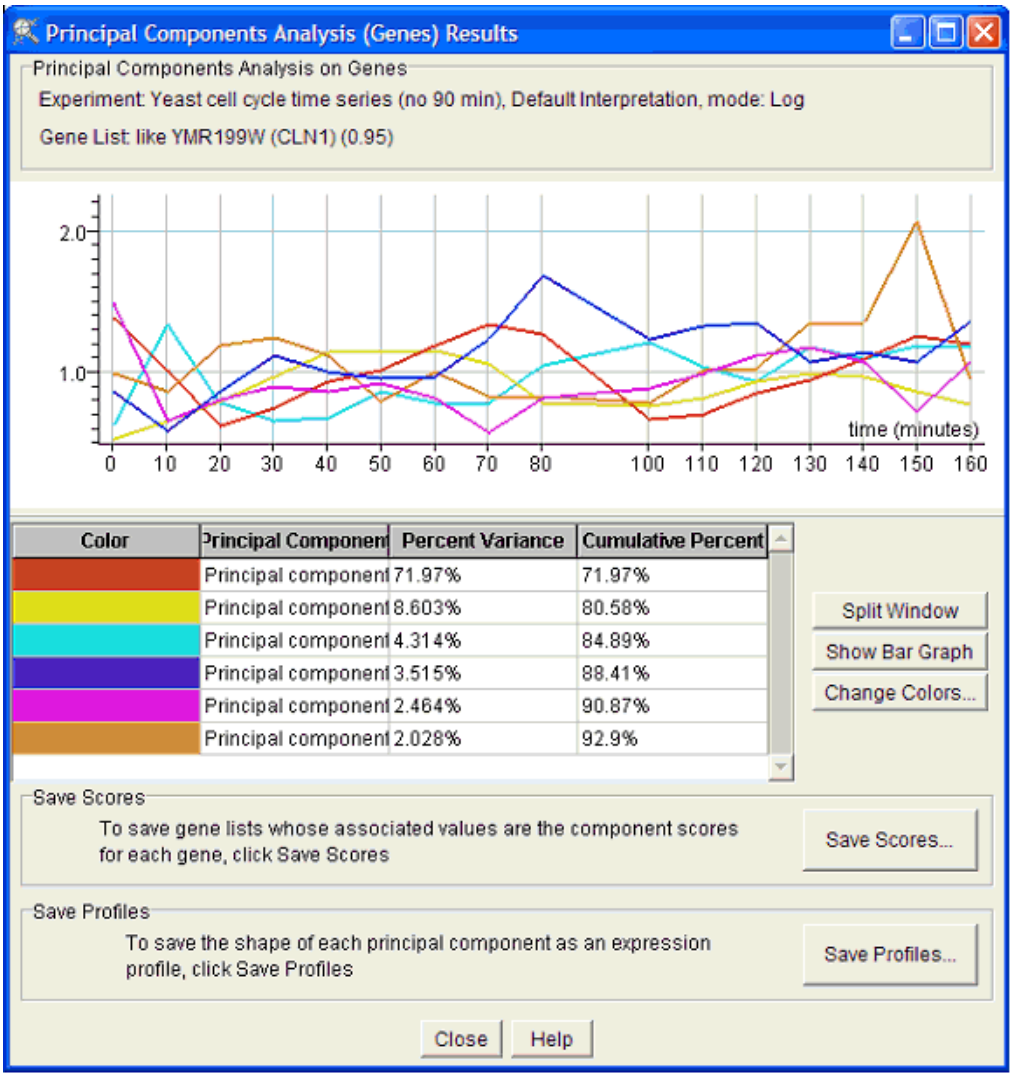

Principal Components Analysis (Genes) Results
Principal Components Analysis on Genes
Experiment: Yeast cell cycle time series (no 90 min), Default Interpretation, mode: Log
Gene List: like YMR199W (CLN1) (0.95)
2.0
1.0
time (minutes)
0 10 20 30 40 50 60 70 80 100 110 120 130 140 150 160
Color
Principal Component
Percent Variance
Cumulative Percent
Principal component 71.97% 71.97%
Principal component 8.603% 80.58%
Principal component 4.314% 84.89%
Principal component 3.515% 88.41%
Principal component 2.464% 90.87%
Principal component 2.028% 92.9%
Split Window
Show Bar Graph
Change Colors...
Save Scores
To save gene lists whose associated values are the component scores for each gene, click Save Scores
Save Scores...
Save Profiles
To save the shape of each principal component as an expression profile, click Save Profiles
Save Profiles...
Close
Help

Figure 15-10 Principal Components Analysis Results Window

To view the PCA on genes results:

1. In the **Results** window, double-click a component.

   The Gene Inspector window opens. Go to "Using the Gene Inspector" on page 7-16 for more information.

   This window shows the eigenvalue and explained variability in the upper-left panel. It contains the following buttons:

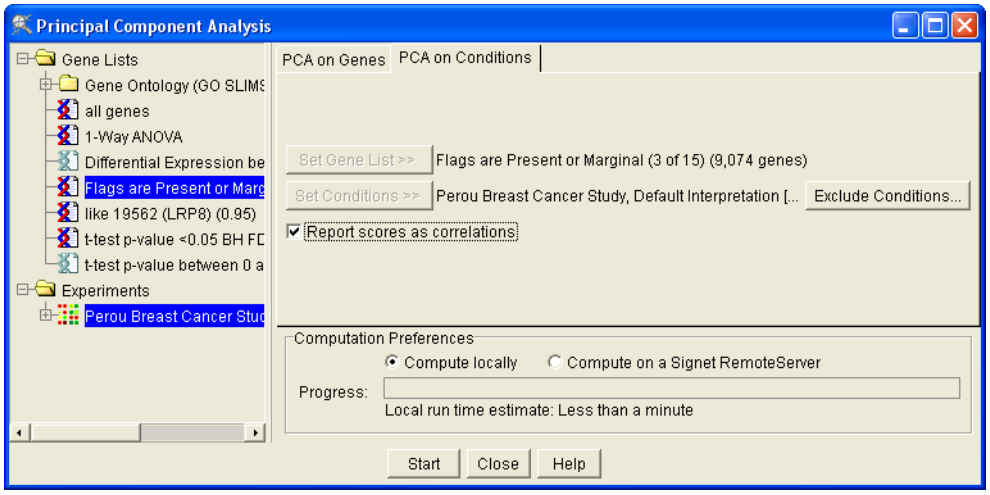
   - **Split/Unsplit Window**—toggles between the default view and splitting the graph by component.
   - **Show Bar/Line Graph**—toggles between the bar and line graph views.
   - **Change Colors**—lets you change the colors used to display the components.
   - **Save Scores—**save gene lists whose associated values are the component scores for each gene.
   - **Save Profiles**—save the shape of each principal component as an expression profile.

2. Click the **OK** button when you are done viewing the results.

Running a PCA on Conditions

To run a PCA on conditions:

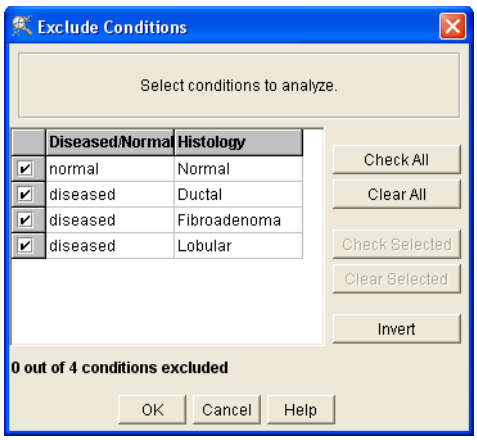
1. Select **Tools > Principal Components Analysis**.
2. Click the **PCA on Conditions** tab.



3. In the **Navigator**, open the Gene Lists folder, select the gene list you want, and then click the **Set Gene List** button.
4. In the **Navigator**, open the Experiments folder, select the experiment you want, and then click the **Set Conditions** button.

5. (Optional) To specify which conditions to exclude from the analysis, click the **Exclude Conditions** button.

   The Exclude Conditions window opens.



   By default, all conditions are selected.

6. Do one of the following:
   - To exclude a condition, clear the check box to its left.
   - To include a condition, select the check box.
   - Click the **Check All** button to include all conditions.
   - Click the **Clear All** button to exclude all conditions.
   - To check or clear a range of conditions, select them in the list and then click the **Check Selected** button or the **Clear Selected** button.
   - Click **OK** button.
7. To specify whether to report scores as correlations or as the values of the principal components for each condition, select or clear the **Report scores as correlations** check box. By default the check box is selected.
8. Specify whether to run the computation locally or on a Signet server.

Viewing PCA on Conditions Results

When the analysis is complete, the PCA Results window opens, displaying each condition as a line in graph mode. The significance of each condition is represented by the color of its graph line, as defined by the Colorbar.

A second window opens that displays a condition scatter plot with the first three components on the axes. For information on using this view, see “Using the Condition Scatter Plot” on page 8-52.

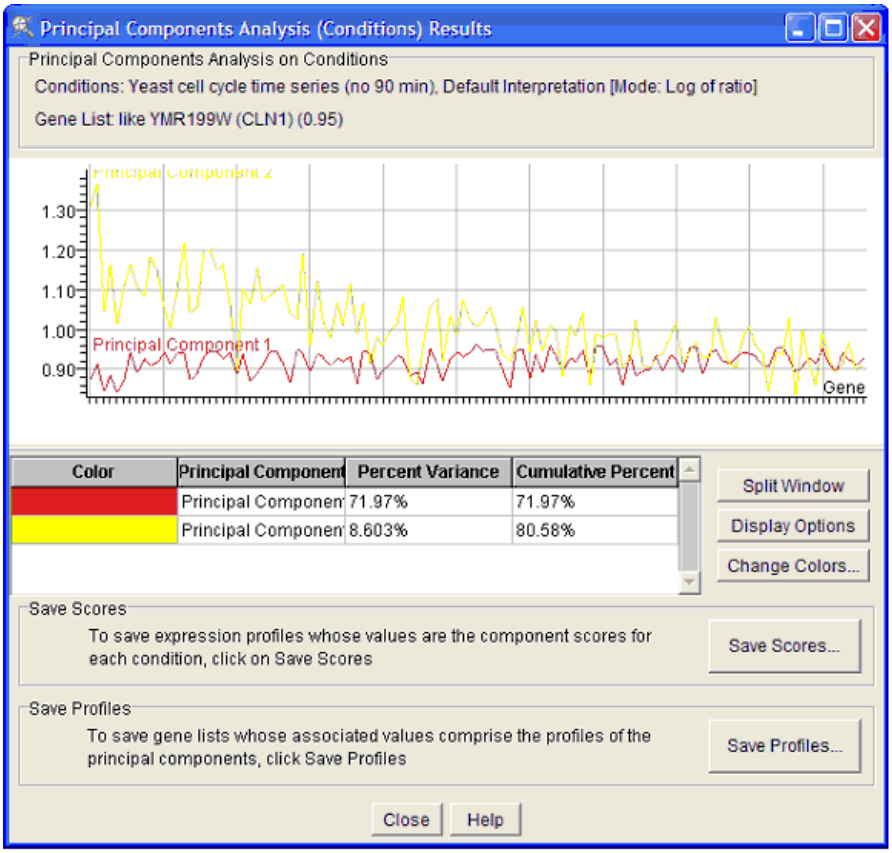


Figure 15-11 Principal Components Analysis Results Window

To view the PCA on conditions results:

1. In the **Results** window, double-click a condition.

   The Gene Inspector window opens.

   Go to "Using the Gene Inspector" on page 7-16 for more information.

   This window shows the eigenvalue and explained variability in the upper-left panel. It contains the following buttons:

   - **Split/Unsplit Window**—toggles between the default view and splitting the graph by component.
   - **Show Bar/Line Graph**—toggles between the bar and line graph views.
   - **Change Colors**—lets you change the colors used to display the components.
   - **Save Scores—**save gene lists whose associated values are the component scores for each gene.
   - **Save Profiles**—save the shape of each principal component as an expression profile. The name of the profile will contain the percent variance that this principle component explains in parentheses.

2. Click the **OK** button when you are done viewing the results.

Interpreting PCA Results

The principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the covariance matrix of the data. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector. The eigenvector of the largest eigenvalue is the first principal component. The eigenvector of the second largest eigenvalue is the second principal component and so on. Principal components which explain significant variability are displayed by GeneSpring in the Principal Components Analysis window.

There are never more principal components than there are conditions in the data.

Viewing Principal Component Loadings in a Scatter Plot

After performing principal components analysis, the Genome Browser displays a 3-D scatter plot in which the loadings for the first, second, and third principal components (representing the largest fraction of the overall variability) are plotted on the X, Y, and Z axes respectively.

In Figure 15-12, each point represents a single gene. Its position on the Y-axis represents the loading of principal component 2. The position on the X-axis represents the loading of principal component 1.

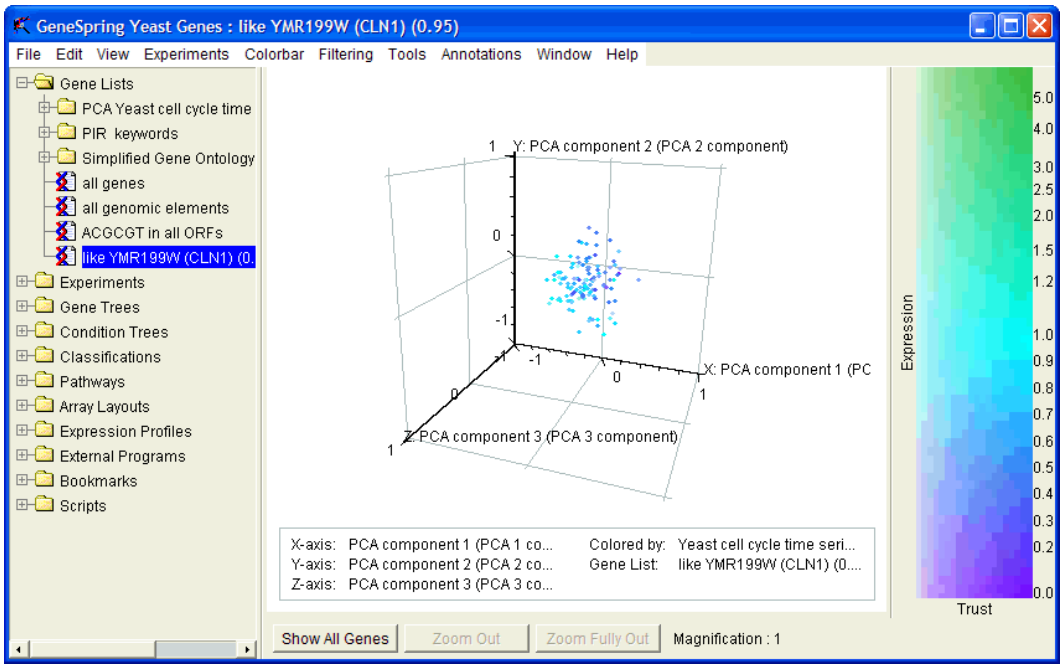


Figure 15-12 PCA Scatter Plot

This view is useful for selecting and making lists of genes that exhibit high levels of one or two principal components. Genes that exhibit high levels of the first principal component and low levels of the second principal component are displayed in the lower right corner of the plot, and genes exhibiting equal levels of the two components lie along the diagonal.

You can change the components that are represented by each axis by right-clicking in the browser and selecting **Display Options**.

Regenerating the PCA Scores Scatter Plot

If you have closed the PCA scatter plot window and have saved the PCA scores as a set of gene lists or expression profiles, you can reproduce the initially displayed scatter plot by doing the following:

To regenerate the PCA scatter plot scores for genes:

1. Do one of the following:
   - Select **View > Scatter Plot**.
   - Select **View > 3D Scatter Plot**.
2. Right-click over the scatter plot and select **Display Options**.
3. In the **Navigator**, select the first gene list you want and assign it to the X axis.
4. Repeat step 3 for the remaining axes.
5. When you are done assigning the gene lists to the axes, click the **OK** button.

To regenerate the PCA scatter plot scores for conditions:

1. Select **View > Condition Scatter Plot**.
2. Right-click over the scatter plot and select **Display Options**.
3. In the **Navigator**, select the first expression profile you want and assign it to the X axis.
4. Repeat step 3 for the remaining axes.
5. When you are done assigning expression profiles to the desired axes, click the **OK** button.

Viewing Principal Components in an Ordered List

The best way to visualize the genes that exhibit the highest levels of an individual component is to use the Ordered List view (Figure 15-13).

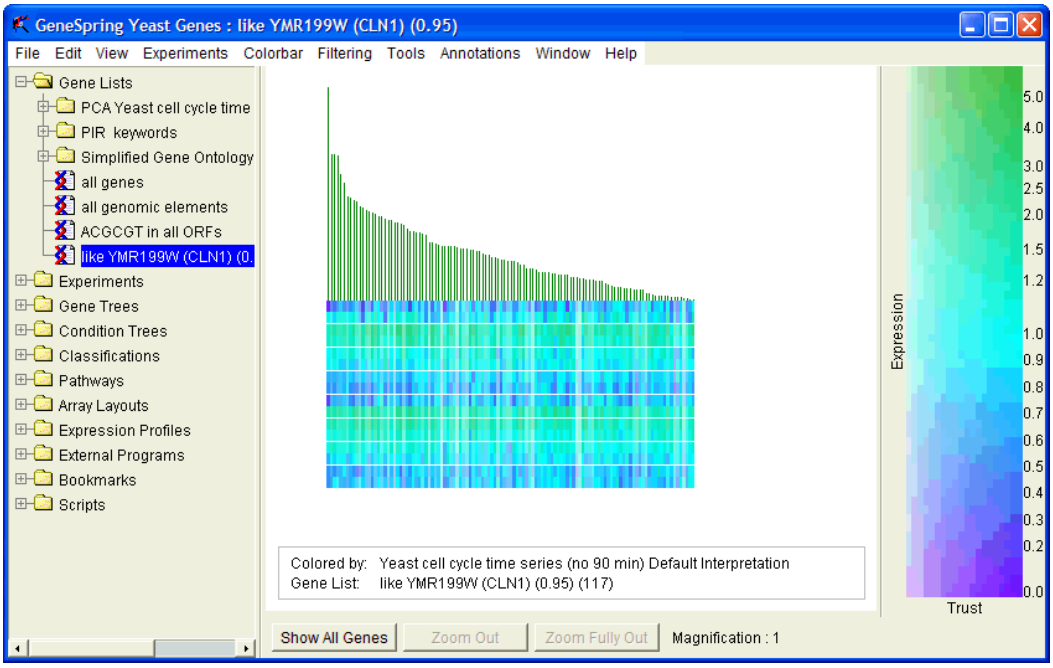


Figure 15-13 PCA Ordered List View

To view PCA in an ordered list:

1. Select **View > Ordered List**.
2. Choose one of the PCA gene lists from the Navigator panel.

   Genes exhibiting the highest levels of the selected principal component are displayed on the left side of the Genome Browser and have the longest lines extending upward from them. For more details, see "Using the Ordered List View" on page 8-38.

References for Principal Components Analysis

Alter O., Brown P.O., Botstein D. *Singular value decomposition for genome-wide expression data processing and modeling.* PNAS 97:10101-6 (2000) http://www.pnas.org/cgi/content/full/97/18/10101.

Cooley, W.W. and Lohnes, P.R. *Multivariate Data Analysis* (John Wiley & Sons, Inc., New York, 1971).

Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations* (John Wiley & Sons, Inc., New York, 1977).

Neal S. Holter et al, Fundamental patterns underlying gene expression profiles: Simplicity from complexity. PNAS 97,8409 (2000) http://www.pnas.org/cgi/content/abstract/97/15/8409.

Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology 24, 417-441, 498-520 (1933).

Kshirsagar, A.M. Multivariate Analysis (Marcel Dekker, Inc., New York, 1972).

Mardia, K.V., Kent, J.T., and Bibby, J.M. Multivariate Analysis (Academic Press, London, 1979).

Morrison, D.F. Multivariate Statistical Methods, Second Edition (McGraw-Hill Book Co., New York, 1976).

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 6(2), 559 -572 (1901).

Rao, C.R. The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhya A 26, 329 –358 (1964).

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pacific Symposium on Biocomputing (2000).

Performing Class Prediction Analysis

Class prediction analysis is designed to predict the value, or “class”, of an individual parameter in an uncharacteristic sample or set of samples. The Class Predictor tool in GeneSpring performs this analysis in two steps. First, the Class Predictor algorithm examines all genes in the training set individually and ranks them on their power to discriminate each class from all the others. Next it uses the most predictive genes to classify the “test set” (for example, the set where the parameter value of interest is unknown).

For example, you could attempt to diagnose the leukemia type of a leukemia patient with the Class Predictor by using expression data from patients whose leukemia type was known. You can also use the Class Predictor simply to find genes whose behavior is related to a given parameter by examining the list of predictor genes.

K-Nearest Neighbors

The class prediction analysis for k-nearest neighbors is designed for experiments with at least 20 or so samples in each class. It is possible to use the Class Predictor when you have very small sample sizes if you disable the p-value cutoff function. For sample sizes of less than 5, specify 1 or 2 number of neighbors and specify 1 in the p-value cutoff field.

To make a prediction, the Class Predictor uses the k-nearest-neighbor method. It selects “k” number of samples near (as measured in Euclidean distance) the unclassified sample, and for each class, computes a p-value that is the likelihood of finding the observed number of this class within the neighborhood members by chance given the proportion of the classes in the training set. The class with the lowest p-value is assigned to the unclassified sample.

You can specify a p-value cutoff, or threshold, such that if there is not sufficient evidence in favor of a particular class, no prediction is made. The p-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest p-value divided by the next lowest p-value.

Support Vector Machines Kernel Function

Support Vector Machines (SVMs) represent a new type of learning system based on recent advances in statistical learning theory. SVMs use pattern recognition and regression estimation to identify sets of genes with a common function from expression data. Using SVMs, you can functionally classify genes using gene expression data from DNA microarray experiments using a set of labeled training data.

Kernel functions perform a non-linear mapping of the original samples into feature space. The SVM technique then finds a separating hyper-plane in this feature space that corresponds to a non-linear decision boundary in the original sample space. The advantage of using a kernel function is that it allows a separating hyper-plane to be found in cases where no linear separation is possible.

SVMs use different similarity metrics, including a simple dot product of gene expression vectors, polynomial versions of the dot product, and a radial basis function. For a given experiment and set of samples, the best kernel function choice is the one which minimizes the misclassification rate.

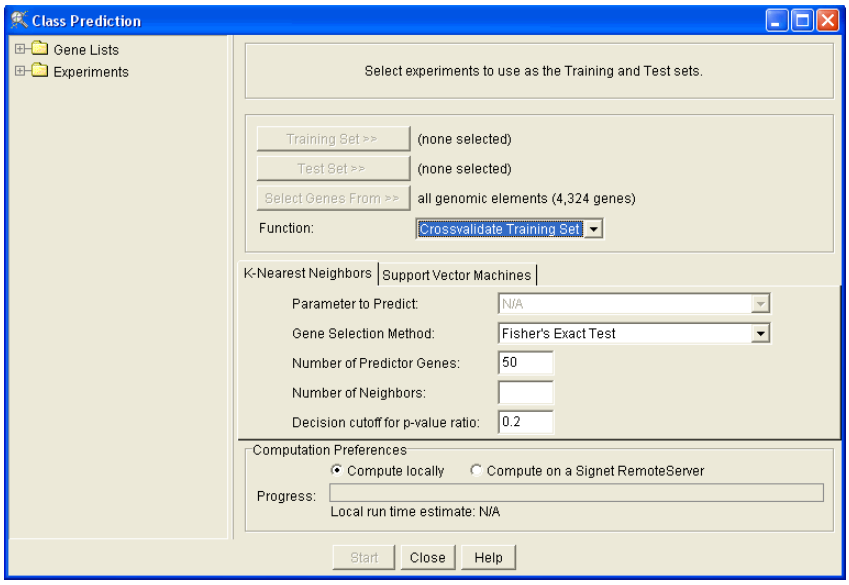
SVMs have many features that make them useful for gene expression analysis, including the ability to manage large data sets, large feature spaces, and outliers.

Performing a Class Prediction Analysis for K-Nearest Neighbors

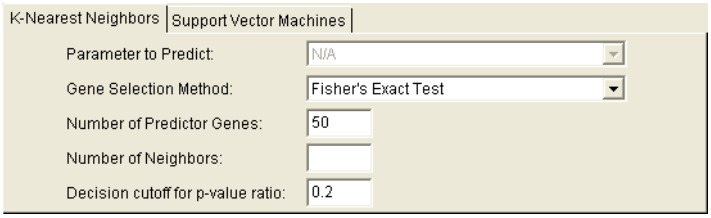
To perform a class prediction analysis for k-nearest neighbors:

1. Select **Tools > Class Prediction**.

   The Class Prediction window opens.



2. Click the **K-Nearest Neighbors** tab.



3. In the **Navigator**, open the **Experiments** folder.
4. Select your training set (the set of samples for which the parameters are already known) and click **Training Set**.
5. Click your test set (the set where the parameter value of interest is unknown) and click **Test Set**.
6. In the **Navigator**, open the **Gene Lists** folder.
7. Select a gene list to be used in the selection process and click **Select Genes From**.
8. In the **Function** list, do one of the following:
   - Select **Predict Test Set** to make a prediction.
   - Select **Crossvalidate Training Set** to evaluate how well the prediction rule can be used to predict the parameter values of the training.
9. Select a parameter in the **Parameter to Predict** box.
10. To control how a subset of genes is selected from the input gene list, select one of the following options in the **Gene Selection Method** list:
    - **Fischer's Exact Test—**Determines if there are nonrandom associations between two categorical variables. In this test, prediction strength is measured by strength of association between sample type and expression level.
    - **Golub Method**—Calculates the difference in means between groups divided by the sum of the standard deviations. The best predictors have large between-group variability and small within-group variability.
    - **All Genes from Selected List**—Indicates that all genes will be used in the analysis; no subset is specified.
11. Specify the **Number of Neighbors**.

    Generally, this number should be no more than half the size of a single class, and no less than 10.
12. Specify a **Decision cutoff for P-value ratio**.

    The p-value cutoff is a threshold such that if there is not sufficient evidence in favor of a particular class, no prediction is made. The p-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest p-value divided by the next lowest P-value.
13. Specify whether to run this process on your local machine or a Signet server.
14. Click the **Start** button.

After the analysis completes, the Prediction Results window opens.

15. Go to "Interpreting Class Prediction Analysis Results" on page 15-32.

Interpreting Class Prediction Analysis Results

The K-nearest Neighbors Prediction Results window (Figure 15-14) opens after you have made a prediction or validated a training set. For convenience, not all of the prediction statistics are visible until you click the **Show Details** button at the bottom of the window. You can also get a description of the analysis by clicking the **Get Text Description** button.

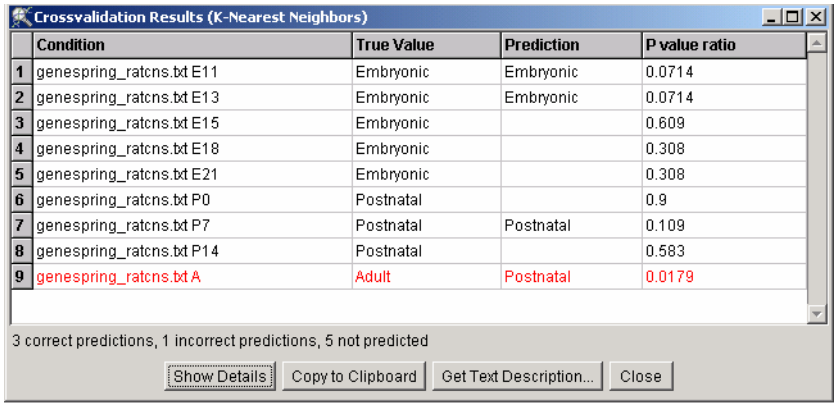

Crossvalidation Results (K-Nearest Neighbors)
Condition
True Value
Prediction
P value ratio
1 genespring_ratcns.txt E11 Embryonic Embryonic 0.0714
2 genespring_ratcns.txt E13 Embryonic Embryonic 0.0714
3 genespring_ratcns.txt E15 Embryonic 0.609
4 genespring_ratcns.txt E18 Embryonic 0.308
5 genespring_ratcns.txt E21 Embryonic 0.308
6 genespring_ratcns.txt P0 Postnatal 0.9
7 genespring_ratcns.txt P7 Postnatal Postnatal 0.109
8 genespring_ratcns.txt P14 Postnatal 0.583
9 genespring_ratcns.txt A Adult Postnatal 0.0179
3 correct predictions, 1 incorrect predictions, 5 not predicted
Show Details
Copy to Clipboard
Get Text Description...
Close

Figure 15-14 K-Nearest Neighbors Prediction Results Window

This window contains the following elements:

- **True Value—**Shows the true value of the class of each sample, as calculated when the parameter for the test set is already known. Compare this with the value in the Prediction column to validate your training set.
- **Prediction—**Shows the predicted class.
- **P-value ratio—**Shows the p-value ratio, or the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest p-value divided by the next lowest p-value.
- **Class counts—**Shows the individual class counts for each sample.
- **P-value—**Shows the probability that individual class counts were found by chance.

If you selected the Crossvalidate Training Set function, the Prediction Results window also displays the prediction rate results. These results include the number classified correctly, the number misclassified, and the number unclassified.

Technical Details for the Class Prediction

This section provides technical details on the class prediction analysis.

Gene Selection

To select genes for use in the predictor, all genes are examined individually and ranked on their power to discriminate each class from all others, using the information on that gene alone. For each gene, and each class, all possible cutoff points on gene expression level for that gene are considered to predict class membership either above or below that cutoff. Genes are scored on the basis of the best prediction point for that class. The score function is the negative natural logarithm of the p-value for a hypergeometric test of predicted versus actual class membership for this class versus all others.

A combined list containing the most discriminating genes for each class is produced as the predictor list. Each class is examined in turn, and the gene with the highest score for that class is added to the list, if it is not already on the list. Then genes with the next highest scores for each class are added. This is continued in rotation among the classes until the specified number of predictor genes is obtained. If you save the list of predictor genes as a Gene List, the best prediction score of the gene among the classes for which it would have been added to the list is saved as the attached number on the list.

Fischer's Exact Test

Fisher's Exact Test looks for an association between expression level and class membership. Each gene is tested for its ability to discriminate between the classes. Genes with the lowest p-values are kept for the subsequent calculations.

In this method, all the measurements for a given gene are ordered according to their normalized expression levels. For each class (parameter value), the predictor places a mark in the list where the relative abundance of the class on one side of the mark is the highest in comparison to the other side of the mark. The genes that are most accurately segregated by these markers are considered to be the most predictive. A list of the most predictive genes is made for each class and an equal number of genes (lowest p-value using Fischer's exact test) are taken from each list.

Golub Method

In the Golub Method, each gene is tested for its ability to discriminate between the classes using a signal-to-noise score, which is given by:

$$\left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right|$$

Where μi and σi, i = 1, 2, are the mean and standard deviation of the expression values over the samples in class i. Genes with the highest scores are kept for subsequent calculations.

Classifying the Test Samples

Based on the selected genes, classifications are then predicted for the independent test data, using the k-nearest-neighbors rule. A sample in the independent set is classified by finding the (user specified) k nearest neighbors of the sample among the training set samples, based on Euclidean distance between the normalized expression ratio profiles of the samples. The class memberships of the neighbors are examined, and the new sample is assigned to the class showing the largest relative proportion among the neighbors after adjusting for the proportion of each class in the training set.

Decision Threshold

P-values are computed for testing the likelihood of seeing at least the observed number of neighborhood members from each class based on the proportion in the whole training set. The class with the smallest p-value is given as the predicted class. The column labeled "P-value ratio" is the ratio of the p-value for the best class to that of the second-best class. The predictor will make a prediction if this ratio is less than the "P-value Cutoff" specified on the initial panel, and will not make a prediction if the ratio is above this cutoff. Setting the p-value cutoff to 1 will force the algorithm to always make a prediction but may result in more actual prediction errors.

References for the Class Predictor

Cover, T.M. and Hart, P.E. (1967) "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, IT-13, 21-27.

Duda, R. O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis, Wiley, New York.

Golub, T.R. et. al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" Science, v286, pp 531-537 (1999).

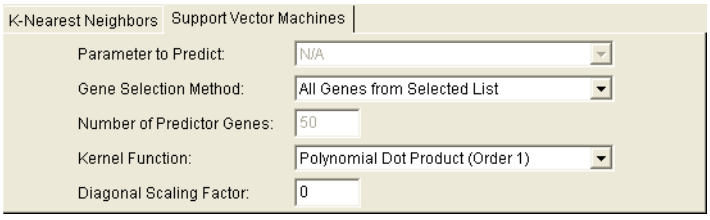
Performing a Support Vector Machines Analysis

To perform an SVM analysis:

1. Select **Tools > Class Prediction**.

   The Class Prediction window opens.

2. Click the **Support Vector Machines** tab.



3. In the **Navigator**, open the **Experiments** folder.

4. Select your training set (the set of samples for which the parameters are already known) and click the **Training Set** button.

5. Click your test set (the set where the parameter value of interest is unknown) and click the **Test Set** button.
6. In the **Navigator**, open the **Gene Lists** folder.
7. Select a gene list to be used in the selection process and click the **Select Genes From** button.
8. In the **Function** list, do one of the following:
   - Select **Predict Test Set** to make a prediction.
   - Select **Crossvalidate Training Set** to evaluate how well the prediction rule can be used to predict the parameter values of the training.
9. Select a parameter in the **Parameter to Predict** box.
10. To control how a subset of genes is selected from the input gene list, select one of the following options in the **Gene Selection Method** list:
    - **Fischer's Exact Test—**Determines if there are nonrandom associations between two categorical variables. In this test, prediction strength is measured by strength of association between sample type and expression level.
    - **Golub Method**—Calculates the difference in means between groups divided by the sum of the standard deviations. The best predictors have large between-group variability and small within-group variability.
    - **All Genes from Selected List**—Indicates that all genes will be used in the analysis; no subset is specified.
11. Specify a **Number of Predictor Genes** to be used in the prediction.

    Note: If you chose **All Genes from Selected List**, this option is not available.
12. Specify a **Kernel Function**. The kernel function acts as a similarity metric between examples in the training set. The options are:
    - **Polynomial Dot Product 1—**The simplest form of similarity metric that provides the dot product between two vectors.

      `k(x,z)=dot(x,z)`, where `dot(x,z)` is the normal vector inner product operator.
    - **Polynomial Dot Product 2, 3, or 4**—a higher-order polynomial kernel. The higher the user-defined parameter, the higher the capacity model that SVM can provide.

      `k(x,z)=(dot(x,z)+1)^d,` where `d` is user defined parameter
    - **Gaussian Kernel—**This kernel has an infinite dimensional feature space and produces models that are good "univeral approximators." As v gets larger, the capacity gets lower.

      `k(x,z)=exp(-(||x-z||^2)/v)`, where `v` is user defined variance.
13. Specify a **Diagonal Scaling Factor**.

    This option is used to control the misclassification rate. It corrects for unbalanced class sizes. A value of 0 assumes that the class sizes are equal.
14. Specify whether to run this process on your local machine or a Signet server.

15. Click the **Start** button.

    After the analysis completes, the Prediction Results window opens.

16. Go to"Interpreting Support Vector Machine Analysis Results" on page 15-36.

Interpreting Support Vector Machine Analysis Results

During the analysis, GeneSpring compute weights for each sample in the training set. Using these weights, a score is calculated for each sample in a test set. Positive scores are assigned to one class, and negative scores are assigned to another class.The scores are then reported as the margins. This corresponds to the distance from the sample to the separating decision boundary.The larger the margin, the farther away a score is from the boundary, and the more confident is the classification.

The SVM Prediction Results window (Figure 15-14) opens after you have made a prediction or validated a training set. For convenience, not all of the prediction statistics are visible until you click the **Show Details** button at the bottom of the window. You can also get a description of the analysis by clicking the **Get Text Description** button.

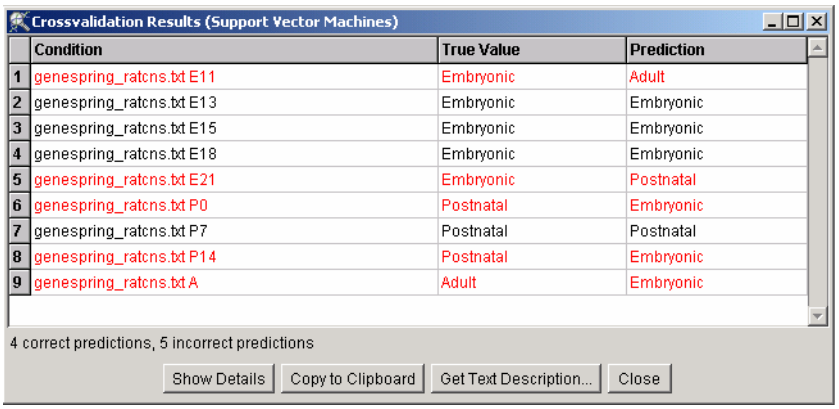


Figure 15-15 SVM Prediction Results Window

This window contains the following elements:

- **True Value—**Shows the true value of the class of each sample, as calculated when the parameter for the test set is already known. Compare this with the value in the Prediction column to validate your training set.
- **Prediction—**Shows the predicted class.
- **P-value ratio—**Shows the p-value ratio, or the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest p-value divided by the next lowest p-value.
- **Class counts—**Shows the individual class counts for each sample.
- **P-value—**Shows the probability that individual class counts were found by chance.

If you selected the Crossvalidate Training Set function, the Prediction Results window also displays the prediction rate results. These results include the number classified correctly, the number misclassified, and the number unclassified.

References for Support Vector Machines

Brown, M. P. S., et. al. (2000) “Support Vector Machine Classification of Microarray Gene Expression Data,” PNAS, v97(1), pp 262-267.

Brown, M. P. S., et. al. (1999) “Knowledge-based analysis of microarray gene expression data by using support vector machines,” UCSC-CRL-99-09.

Furey, T.S., et. al. (2000) “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” Bioinfomatics, v16, pp 906-914.

Jaakkolay, T., et. al. (1999) “A discriminative framework for detecting remote protein homologies,” MIT Artificial Intelligence Laboratory, Department of Computer Science, pp 1-28.

Finding Significant Parameters

To identify genes that are correlated or uncorrelated with experimental parameters or sample attributes, GeneSpring 7 now includes new options for comparing expression data to a single numeric or non-numeric attribute or parameter. Three options are available:

- Find Significant Parameters Using ANOVA
- Find Significant Parameters using an Association Test
- Find Significant Parameters using a Correlation to Parameters

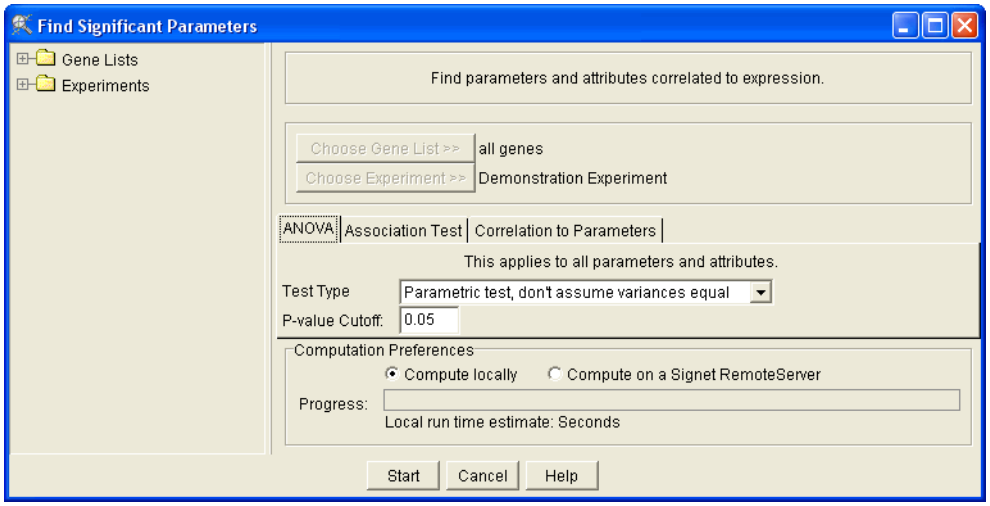
Finding Significant Parameters Using ANOVA

The Find Significant Parameters using ANOVA option performs an ANOVA for each gene in the specified starting list, over all parameters and attributes. Both numeric and non-numeric parameters or attributes can be tested; however, only parameters/attributes that define replicate samples are included. Use this option when you have replicate samples for each parameter (or attribute) in which you are interested, and want to perform a statistical test using the variability within and between replicate groups.

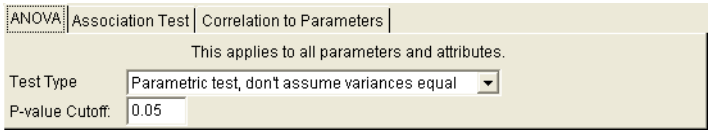
To find significant parameters using ANOVA:

1. Select **Tools > Find Significant Parameters**

   The Find Significant Parameters window opens.



2. Click the **ANOVA** tab.



3. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

4. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

   In the **Test Type** list, select the type of test to perform. There are four testing options:

   - **Parametric test, assume variances equal**—Filter based on the results of a Student's two-sample t-test for two groups or a one-way analysis of variance (ANOVA) for multiple groups.
   - **Parametric test, don't assume variances equal**—Filter based on the results of an ANOVA or Welch's approximate t-test for two groups. This is the most appropriate test for standard experiments in which the global error model is not turned on or should not be used in the analysis.

- **Parametric test, use all available error estimates**—Filter based on the variances estimated by the Cross-Gene Error Model. If the Cross-Gene Error Model is not turned on, this test is equivalent to the Parametric test.
- **Non-Parametric test**—Filter based on the rank of each sample, rather than the expression level. Non-parametric comparisons use the Wilcoxon two-sample rank test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups. This test is most successful if you have more than five replicate samples in each group.

5. Select a **P-value cutoff** for genes that pass the filter.

   The p-value indicates a probability, with a value ranging from zero to one, that the difference observed between groups is due to chance. The lower the p-value, the more significant the difference between the groups.

6. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

7. Click the **Start** button.

   As the ANOVA executes, you can watch its progress in the Progress bar. Results are shown in the Results window.

8. Go to "Interpreting Finding Significant Parameters Results" on page 15-43 for more information.

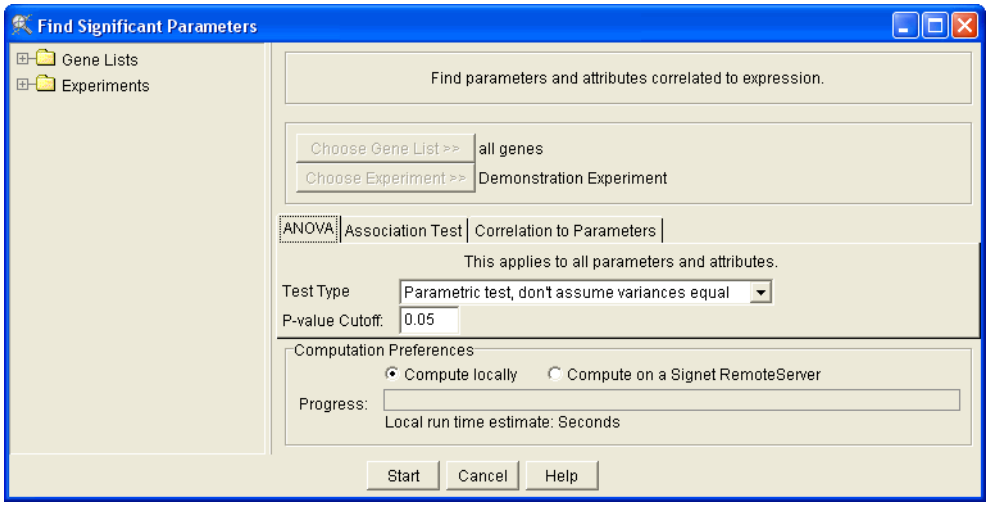
Finding Significant Parameters Using an Association Test

The Find Significant Parameters using an Association Test option performs an association test for each gene, over all parameters and attributes. Both numeric and non-numeric parameters and attributes can be tested. This test works best, however, for parameters or attributes that define a small number of groups that have a relatively large number of samples in each group. The specific test used is Fisher's Exact Test for association between expression level and class membership.

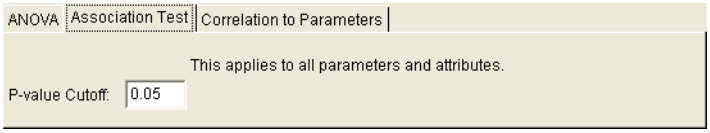
To find significant parameters using an association test:

1. Select **Tools > Find Significant Parameters**

   The Find Significant Parameters window opens.



2. Click the **Association Test** tab.



3. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

4. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

5. Select a **P-value cutoff** for genes that pass the filter.

   The p-value indicates a probability, with a value ranging from zero to one, that the difference observed between groups is due to chance. The lower the p-value, the more significant the difference between the groups.

6. In the **Computation Preferences** box, specify whether to run the analysis locally or on a remote execution server.

   This option lets you run the computation on a Signet server. To use this option, you must be connected to Signet.

7. Click the **Start** button.

As the analysis executes, you can watch its progress in the Progress bar. Results are shown in the Results window.

8. Go to "Interpreting Finding Significant Parameters Results" on page 15-43 for more information.

Finding Significant Parameters Using Correlation to Parameters

The Find Significant Parameters using a Correlation to Parameters option calculates the correlation between each gene and the values of each numeric parameter or attribute. Non-numeric parameters or attributes cannot be tested using this method. The correlation coefficient is calculated using the parameter (or attribute) values, and the expression values for each gene. If replicate samples exist, they will be averaged. This test emphasizes linear relationships between average expression level and the level of the parameter or attribute.

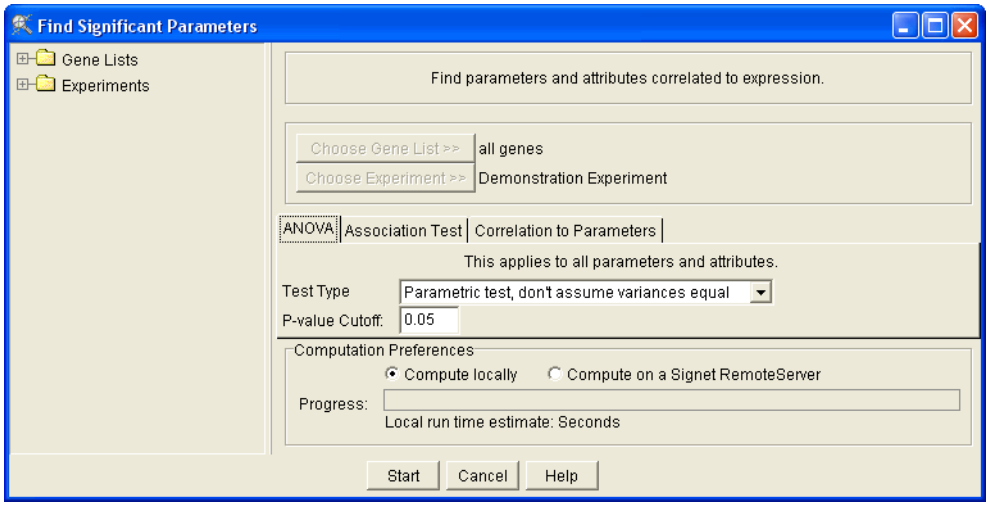
The attributes that you choose should all be on the same scale of measurement. For example, all the Age attribute values should start counting from the same point, rather then comparing Age from conception numbers to Age from birth numbers.

If an attribute has non-numeric values for some samples, those samples will be treated as if they have no values.

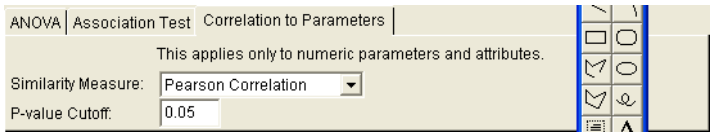
To find significant parameters using a correlation to parameters:

1. Select **Tools > Find Significant Parameters**

   The Find Significant Parameters window opens.



2. Click the **Correlation to Parameters** tab.



3. In the **Navigator**, select the gene list containing the set of genes you would like to analyze, and then click the **Choose Gene List** button.

   This option lets you select the gene list that contains the set of genes you would like to analyze. Statistical tests will be performed only on genes in the selected gene list. It is recommended that the **All Genes List** option not be used. Instead, use a list of genes that has been filtered to remove genes with measurements mostly in the noise range or mostly flagged Absent.

4. In the **Navigator**, select the experiment interpretation you want to analyze, and then click the **Choose Experiment** button.

   This option lets you select the experiment and its proper interpretation to analyze. If you are using parametric tests, then your experiment interpretation should be in log-of-ratio mode.

5. In the **Similarity Measures** list, select the type of correlation you want. The options are:

   - Standard Correlation
   - Smooth Correlation
   - Change Correlation
   - Upregulated Correlation
   - Pearson Correlation
   - Spearman Correlation

   For more information on measures of similarity, see “Similarity Measures” on page 15-20.

6. Select a **P-value cutoff** for genes that pass the filter.

   The p-value indicates a probability, with a value ranging from zero to one, that the difference observed between groups is due to chance. The lower the p-value, the more significant the difference between the groups.

7. Click the **Start** button.

   As the analysis executes, you can watch its progress in the Progress bar. Results are shown in the Results window.

8. Go to “Interpreting Finding Significant Parameters Results” on page 15-43 for more information.

Interpreting Finding Significant Parameters Results

Results of the Find Significant Parameters analysis are displayed in the Find Significant Parameters Results window. This window looks at every attribute and parameter, and returns a list of genes that are similar to each one. It contains two tabs: the Parameters tab (Figure 15-16) and the Genes tab (Figure 15-17).

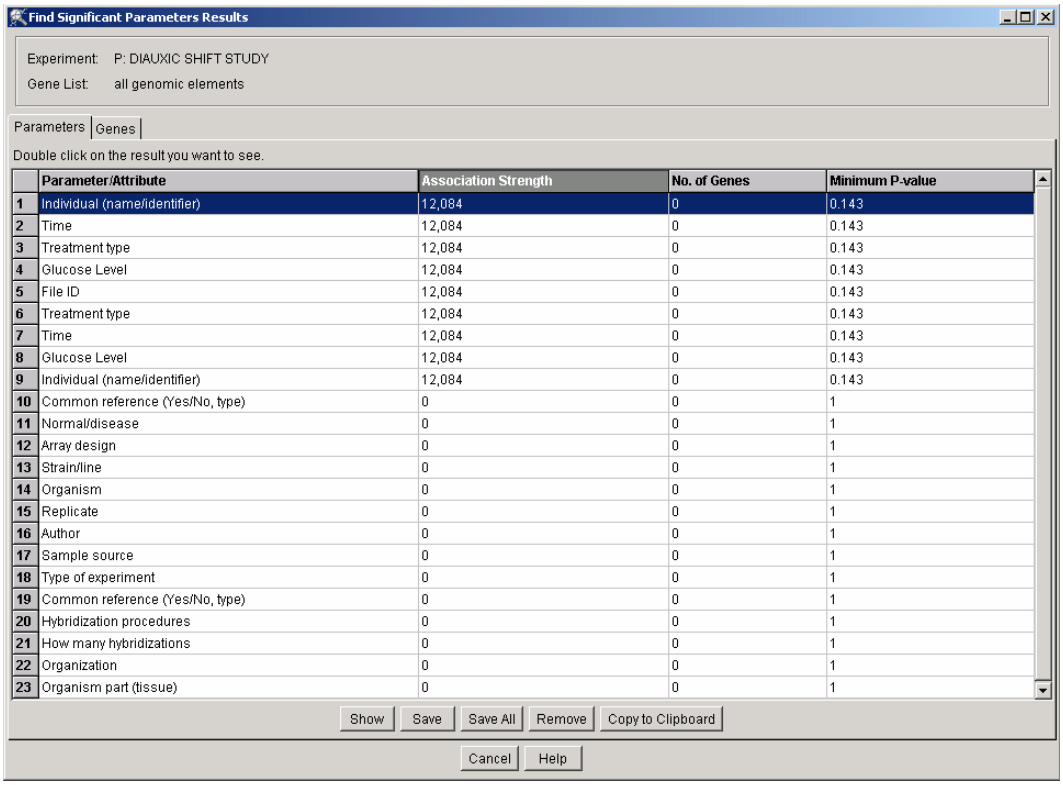
Parameters Tab

The Parameters tab displays the results of the Find Significant Parameters analysis. Each row in the table summarizes a gene list. You can inspect and save a single gene list, or all gene lists in the table.

Table

The table in the Parameters tab displays one row for each parameter or attribute tested.The table is initially sorted by **Association Strength**. Clicking on a column header sorts the table by that column.

The **Association Strength** column records the sum of the negative logs of the p-values for the gene list in question. The **No. of Genes** column says how many genes passed the cutoff for that parameter. The **Minimum P-Value** gives the best p-value for any of the genes associated with the parameter or attribute.



	Parameter/Attribute	Association Strength	No. of Genes	Minimum P-value
1	Individual (name/identifier)	12,084	0	0.143
2	Time	12,084	0	0.143
3	Treatment type	12,084	0	0.143
4	Glucose Level	12,084	0	0.143
5	File ID	12,084	0	0.143
6	Treatment type	12,084	0	0.143
7	Time	12,084	0	0.143
8	Glucose Level	12,084	0	0.143
9	Individual (name/identifier)	12,084	0	0.143
10	Common reference (Yes/No, type)	0	0	1
11	Normal/disease	0	0	1
12	Array design	0	0	1
13	Strain/line	0	0	1
14	Organism	0	0	1
15	Replicate	0	0	1
16	Author	0	0	1
17	Sample source	0	0	1
18	Type of experiment	0	0	1
19	Common reference (Yes/No, type)	0	0	1
20	Hybridization procedures	0	0	1
21	How many hybridizations	0	0	1
22	Organization	0	0	1
23	Organism part (tissue)	0	0	1

Figure 15-16 Find Significant Parameters Results: Parameters Tab

Buttons

- **Show**—Displays the **Gene List Inspector** for the selected row.
- **Save**—Displays the **Save Gene List** window for the selected row.
- **Save All**—Displays **Save Gene List Folder** window containing all of the gene lists in the table. The default name for the folder contains the name of the experiment used in the **Find Significant Parameters** window. Each gene list has an associated p-value.
- **Remove**—Removes the selected gene list from the table.

Genes Tab

The table in the Genes tab contains one row for each gene in the gene list used in the analysis.The table rows contain genes and the table columns contain attributes. The genes can be sorted by clicking on any of the column headers. The Colorbar is graphed on a log scale

Buttons

- **Copy to Clipboard**—Copies the entire table to the clipboard, not just the selected portion.
- **Change Colors**—Displays a window for changing colors in the view.

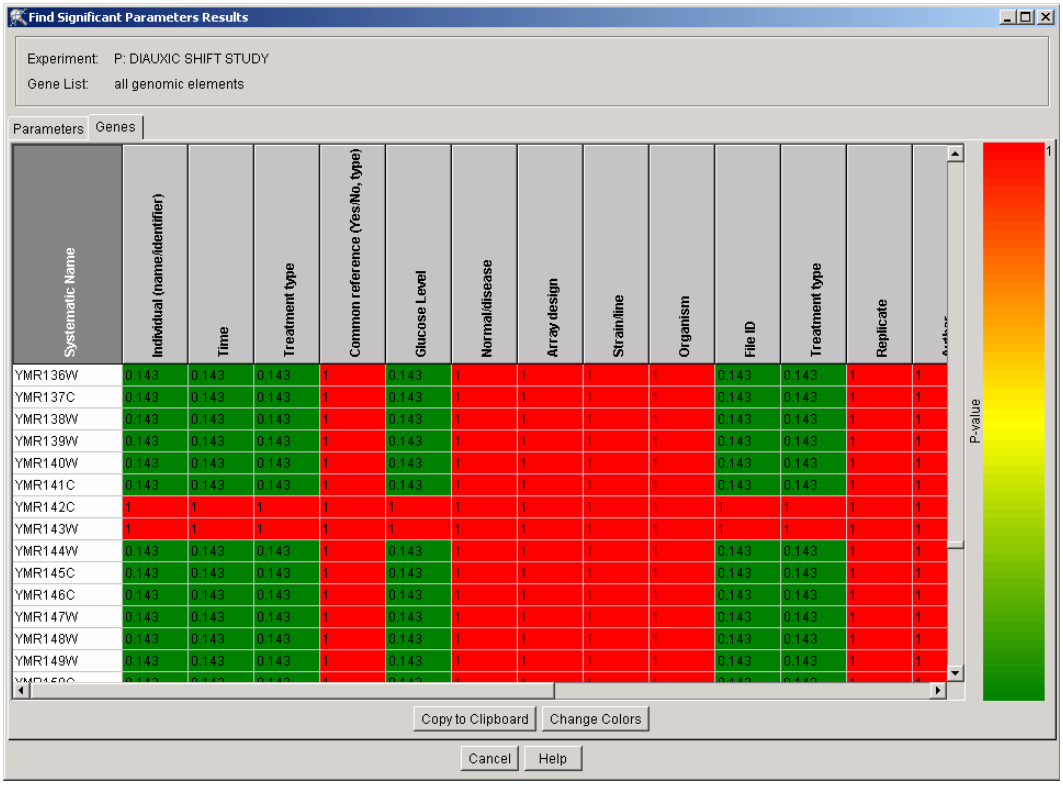


Figure 15-17 Find Significant Parameters Results: Genes Tab

Technical Details for the Find Significant Parameters

This section describes the test statistics used to calculate the ANOVA, association test, and correlation to parameters analysis for the Find Significant Parameters option.

ANOVA

An ANOVA is performed across the groups defined by the distinct values of the parameter or attribute. Multiple test corrections are not used. The Association Strength is calculated by:

$$\sum (-\log p_i)$$

where p_i's are the p-values from the ANOVA calculation, and the sum is taken over all genes in the specified starting gene list.

Association Test

The calculation here is essentially the same that is used in testing genes for their predictive power in the Predict Parameter Values tool. The calculation follows the following steps for each gene:

For each value of the parameter or attribute, the following calculations were performed:

- Perform a series of Fischer's exact tests for association between the number of samples corresponding to the given value of the parameter or attribute and expression level. Use cut-points given by the expression levels of the gene in the samples corresponding to the parameter value.
- Keep the lowest p-value from the above series of tests.
- Keep the lowest p-value from each iteration of step 1 (across all parameter values)
- Assign this p-value to the gene.

The number of genes passing the p-value cutoff is reported, along with the lowest p-value across all genes in the starting gene list. The Association Strength is calculated using the same test statistic as for the ANOVA. Go to "ANOVA" on page 15-45 for more information.

Correlation to Parameters

Correlations between parameter values and gene expression values are calculated using the same test statistics as in the Filter on Parameter window. Go to "Filtering on Parameters" on page 9-16 for more information. The Association Strength is calculated using the same test statistic as for the ANOVA. Go to "ANOVA" on page 15-45 for more information.

Converting Correlations to P-values

To test the null hypothesis H_0: $\rho = 0$ versus the 2-sided alternative (where ρ is the "true" population correlation), the following test statistic is used:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

The derived value is compares to the *t* distribution, where *r* is the computed correlation and *n* is the length of the vectors (number of samples or conditions).

The degrees of freedom used is n – 2 for smooth, Pearson and Spearman correlations; and n – 3 for change and upregulated correlations.

The p-value is:

$$2\Pr(t_{n-2} > |t|)$$

16

Using Scripts, External Programs, and Plugins

This chapter explains how to use scripts, external programs, and plugins with GeneSpring. It covers the following topics:

- Working with Scripts
- Inspecting Scripts
- Using the Script Editor
- Working with Building Blocks
- Managing Scripts
- Getting Help with Scripts
- Working with External Programs
- Working with Plugins

Working with Scripts

You can create custom scripts to automate repetitive analytical tasks, ensure consistency in the analysis process, and simplify data analysis management. You can design scripts that automatically upload results to Signet, or combine scripts with basic functions to perform more complex analyses.

Scripts are re-usable and can be applied to any data set. The analysis steps can be recorded and saved for future use and for sharing with colleagues. Scripts are also ideal for any high-throughput environment.

GeneSpring also provides BioScript Library—a ready-made collection of scripts that span the complete analysis process. It includes scripts for automating a broad range of quality control, statistical analysis, and biological data query tasks.

You can create your own scripts using the Silicon Genetics Script Editor or use a pre-defined script provided by GeneSpring. This section covers the following topics:

- Scripts Folder
- Pre-defined Scripts
- Configuring Script Preferences
- Running Scripts Locally

- Running Scripts Remotely
- Handling Scripts that Generate Long Names or Invalid Characters
- Uploading Scripts to Signet

Scripts Folder

All scripts, including complimentary scripts shipped with GeneSpring, are stored in the Scripts folder.

Pre-defined Scripts

GeneSpring provides the following pre-defined scripts for you to use.

- **2-fold Expression Change—**This script makes a gene list of all genes in a selected experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition.
- **2-fold Expression Change AND Filter on Noise NOT Input**—This script combines the 2-fold Expression Change List and Filter on Noise scripts to produce a single gene list that passes both filters, but does not have any genes on the input gene list.
- **Best k-means—**Given an experiment and a gene list, this script creates four k-means classifications with three, five, eight and 15 clusters respectively and selects the classification with the highest explained variability. The selected k-means appears in a results window.
- **Clustering 2-fold Change List—**This script creates a gene list of all genes in a selected experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition and then creates a gene tree, an condition tree, a k-means classification, and a self-organizing map.
- **Filter on Noise—**Creates a list of genes that have control strengths equal to or greater than a user-supplied cutoff in at least half of the conditions in the experiment.
- **Find List of Similar Genes**—This script makes a gene list for each of the genes in a selected experiment if there are at least five genes with similar expression profiles.
- **Pairwise Comparison—**Returns a list of genes that are two-fold overexpressed in at least one condition in an experiment at compared with a specified experiment.
- **Probe Entire Enterprise Repository for Similar Conditions (PEER-C)**—Given a condition, this script searches through Signet for conditions similar to the input condition. If the same sample is normalized differently in different experiments, both normalizations are compared.
- **Select k-means—**Given an experiment and a gene list, this script creates two k-means classifications with the numbers of clusters specified by the use and chooses the k-means cluster with the highest explained variability as the result.
- **Send Clustering Results to Signet—**This script creates a gene tree, condition tree, k-means classification, and self-organizing map using a list of all genes in an experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition and automatically sends the results to Signet.

- **Series of k-means (increments of 5)—**Generates 10 k-means classifications, each with a differing number of starting sets and returns the classification with the highest explained variability.

Configuring Script Preferences

The Computation tab (Figure 16-1) lets you specify settings for running scripts.

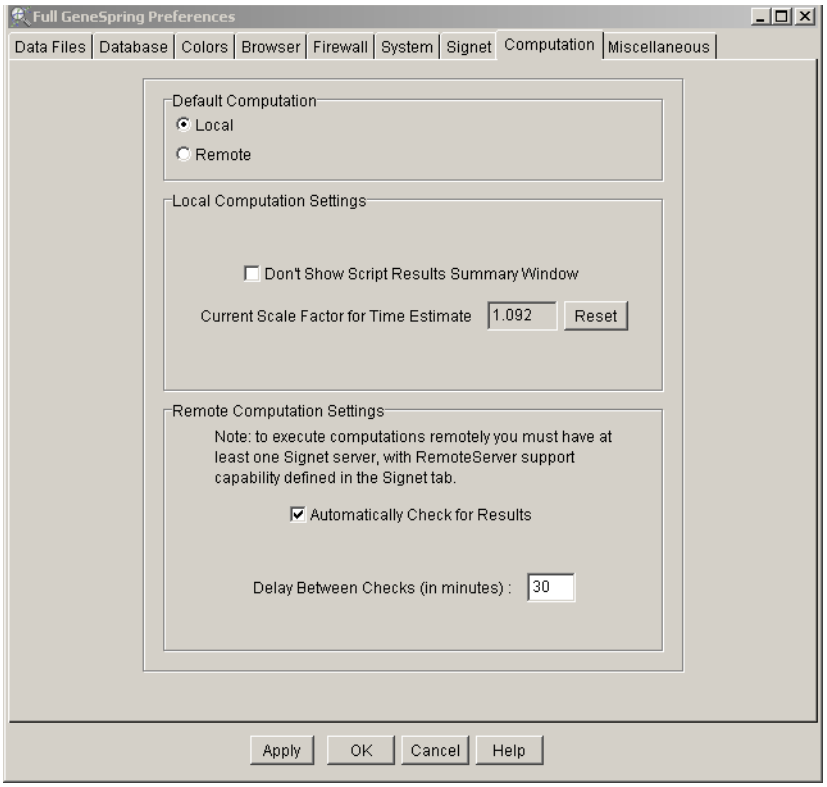


Figure 16-1 Preferences Window: Computation Tab

This window provides the following options:

- **Default Computation**—Select **Local** to have scripts run on your local machine by default. Select Remote to run scripts on a remote execution server by default.
- **Local Computation Settings**—Select **Don't Show Script Result Summary Window** to skip the Script Result Summary when a script completes its execution. This option applies to scripts that produce fairly complex multiple group results.The **Current Scale Factor for Time Estimate** option lets you reset the multiplier for GeneSpring's internal estimate of how long an analysis will take back to the default value of 1. This can be useful if you have significantly changed hardware or settings on your computer (added more RAM, etc.).
- **Remote Computation Settings—**Select **Automatically check for results** to have GeneSpring automatically check whether your script has finished running. Specify how often to check by entering a number of minutes in the **Delay between checks** box.

Running Scripts Locally

Scripts are time-saving tools allowing a long series of data analysis steps to be performed at once. This section explains how to run scripts locally on GeneSpring.

Run Script Window

The Run Script window (Figure 16-2) lets you execute scripts and view information about them. Depending on which script you choose the run, the options that appear in this window will differ. If you have a connection to Signet and are using remote execution servers, you can execute the script on a remote server.

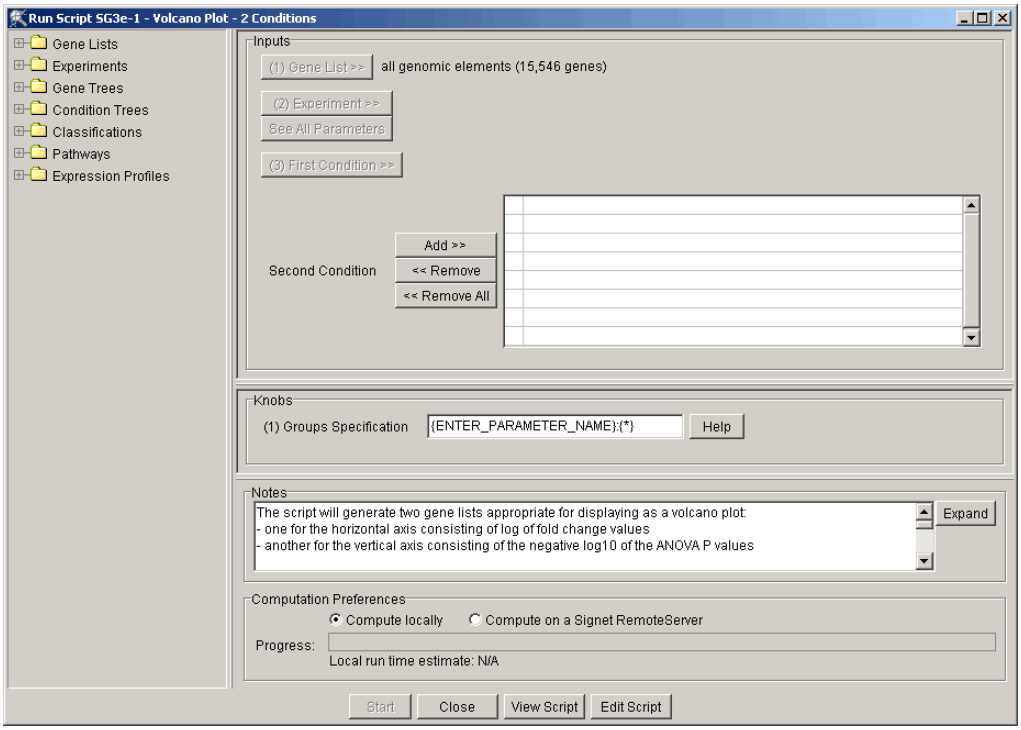


Figure 16-2 Run Script Window

This section describes elements that can appear in the Run Script window.

Navigator

The Navigator contains a subset of folders and files from the main GeneSpring window that are applicable to the script inputs.

Inputs

The Inputs box displays the characteristics of the input data that will be executed by the script. Depending on the type of input you specify, the characteristics you can specify will vary and the buttons that appear in this window will change accordingly.

Knobs

A knob enables you to specify additional parameters related to the script. A knob is value entered when running the script. If a knob has been specified for this script, the appropriate way to set its value appears in the window. If not, the message “this script has no knobs” displays instead.

Notes

The Notes field displays notes added to the script. To enter or edit notes, you need to use the ScriptEditor.

Computation Preferences

The Computation Preferences section lets you specify how the script is to be run. The options are Compute Locally or Compute on a Signet server. To run a script on a remote server, you must be connected to Signet.

Progress

The Progress bar depicts the progress of the script’s execution. An estimate of the time required to run the script also appears in this part of the window. The estimates can include the following:

- Seconds
- Less than a minute
- A few minutes
- Less than an hour
- A few hours
- Many hours

The Progress function cannot determine the running time of external programs and script plugins. Instead, the running time will use a value of “0”, which translates to seconds.

Buttons

- **Start**—Executes the script, as follows:
  - If the script is executed locally, the Start function executes the script and presents the results so they can be saved locally. Go to “Running Scripts Locally” on page 16-4 for more information.
  - If the script is executed remotely, the Start function sends it to an execution queue where it waits for the next available remote server. Following script execution, the remote server sends the results to GeneSpring via Signet where they can be retrieved. Go to “Running Scripts Remotely” on page 16-9 for more information.
- **Close**—Closes the Run Script window.
- **View Script**—At the bottom of the window is a button labeled **View Script**. Click this button to open a new window containing a graphical representation of the script. On this window, click the **Details** button to view detailed information in the Script Inspector. For more information on the Script Inspector, see “Script Inspector Window” on page 16-13.

- **Edit Script**—Displays the ScriptEditor for the script.
- **Expand**—Displays the contents of the **Notes** box in a separate window.

Opening the Run Script Window

To open the Run Script window:

- To open the Run Script window, do one of the following:
  - In the **Navigator**, open the **Scripts** folder, and click the script you want.
  - In the **Navigator**, open the **Scripts** folder, right-click the script you want, and then select the **Run** command.
- To run the script you are currently editing, open the **Script Editor** and select **File > Run** in the Genome menu.

Running Scripts Locally

To run a script locally:

1. In the **Navigator**, select the **Scripts** folder.
2. Select one of the demo scripts.

   The Run Script window opens, displaying various elements of the selected script.
3. In the **Navigator**, select the data object you want, and click the appropriate button in the **Inputs** box.
4. Change the values for any knobs you want.

   The Input fields do not recognize numbers with spaces or commas. Use periods (.) as your decimal markers.

   If your script requires an array of numbers (for example, the weights associated with a complex correlation), a table appears in which you can enter these numbers. Enter one number per line. You can also paste numbers into this table in tab-delimited format from a text or excel file. The order of these numbers must match the order of the inputs that they describe.
5. Specify whether to conduct the execute the script locally or on a remote execution server.

   Note: You must be logged in to Signet to execute a script remotely.
6. Click the **Start** button.

   This button is not active until all required data has been entered.
   - If you are executing locally, the script begins to execute.
   - If you are executing remotely, the script is sent to an available execution server where it is either executed immediately or placed in a queue.

- If your script is supposed to upload data to a regulatory compliant Signet, you will be prompted to provide an electronic signature.
- If your script produces results that generate a Name Problems window, go to "Handling Scripts that Generate Long Names or Invalid Characters" on page 16-12 for more information.

7. If your script returns a data object, do one of the following:
   - To save your result, enter a **Name**, add any comments in the **Notes** field, and then click the **Save** button.
   - If you do not want to save the results, click the **Cancel** button or close the window.

Specifying Complex Parameters

Some of the sample scripts included with GeneSpring require you to enter parameters for data file restriction. Which data file format to search, and which columns in that data file format that are to be searched are entered as special text strings in the Filter Columns Specification knob.

Guidelines

Use curly brackets to indicate the file format. Within the curly brackets give the column number or the column header of the columns to search. The order of the curly brackets must match the order of the tabs given in GeneSpring's interface.

The first tab is the format of the majority of samples in the experiment; the second tab defines the second-most common format for that experiment, etc. If the formats describe the same number of samples, look at the GeneSpring GUI to determine the order of the formats. For example:

```
{Column Number or Column Name1; Column Number or Column Name2}{}{}
```

```
{1;3}{}{"identifier"}
```

Group Specification Text String

The groups of conditions to compare (defined by the parameter values for specified parameters) are entered as a special text string in the Group Specification knob of the Run Script window.

General Format

To write an ANOVA text string, you use curly brackets to surround the names of all the experimental parameters in which you are interested in comparing. Use semicolons to separate the parameter names. If you only want to compare specific levels of the different parameters, enter the names of the specific levels followed by a colon. Then use curly brackets to surround each group of parameter values you want to compare. For example:.

```
{Parameter Name1; Parameter Name2} : {Parameter1Value1;
Parameter2Value1}{Parameter1Value2; Parameter2Value2}
{Treatment; Gender} : {Control; Male}{Control; Female}
```

Compare these statements to the Select Groups Manually window (Figure 16-3) using the following experimental design:

```
{Leukemia type;Gleason Score} : {ALL;8} {AML;8}
```

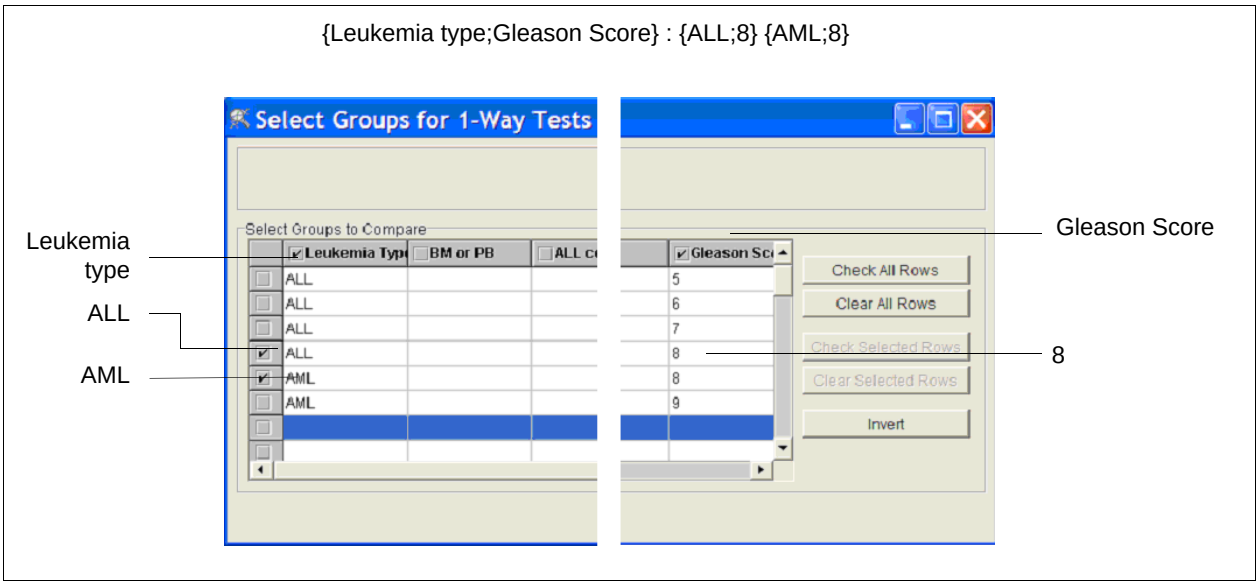


Figure 16-3 Statistical Analysis (ANOVA): Select Groups to Compare

Comparing All Parameter Values

If you want to compare all levels of each parameter, you can write a long string defining all the different levels. However, GeneSpring has provided alternative methods for comparing parameter values.

- If you only state the parameters and do not include the colon and parameter values, then GeneSpring will assume that every different combination should be tested.
- An asterisk (*) indicates that every level of a given parameter value. Thus, if you includes the colon, but replace the parameter values with an asterisk (*), then GeneSpring will compare every different level.

For example if there are two parameters, each with two parameter values the string might look like the following:

```
{Treatment; Gender} : {Control; Male}{Control; Female}{Treated;
Male}{Treated; Female}
```

```
{Treatment; Gender}
```

```
{Treatment; Gender} : {*; *}
```

Comparing All Levels Of One Parameter

If you want to compare all levels of a single parameter while keeping another parameter value constant, you can write out a long string defining all the different levels. However, GeneSpring has provided an alternative method for comparing all levels of one parameter. An asterisk (*) is a short cut that represents every level of the parameter's value. Thus, if the "Treatment" parameter has only two levels, the following strings are equivalent:

```
{Treatment; Gender} : {Control; Male}{Treated; Male}
{Treatment; Gender} : {*; Male}
```

Punctuation

Whenever you use one of the following characters in parameter names or parameter values, you must precede it with a backslash:

```
{ } * ; : \
```

Refer to Table 16-1for the correct usage.

Table 16-1 Group Specification Text String Punctuation

Character used in name	Characters to use in text string
\	\\
{	\{
}	\}
*	*
;	\;

For example, given the parameter name "Control\Treatment" with values "Control*" and "{Treatment}", the statement would be written as:

```
{Control\\Treatment; Gender} : {Control\*; Male}{\{Treatment\};
Male}
```

Running Scripts Remotely

You can execute a script on a remote execution server if the following conditions are met:

- You have a valid user name and password for Signet
- Signet is enabled with remote execution servers
- At least one remote server is installed and configured

The results are saved on Signet and returned to GeneSpring when the script completes. They can be retrieved in GeneSpring upon your request.

Using a remote execution server is highly recommended when you are performing time-consuming tasks such as clustering very large data sets. The remote execution server can remove the burden on your local computer while it completes the necessary computation.

Remote Execution Queue Window

The Remote Execution Queue window (Figure 16-4) shows the status of your running, paused, pending, or completed jobs. If you an administrator, you can view the status of all jobs in the queue.

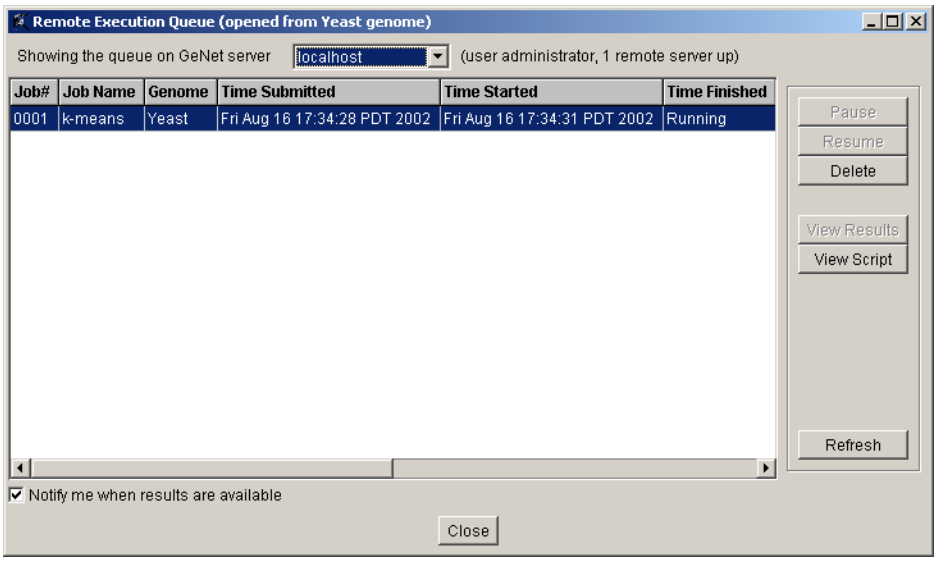


Figure 16-4 Remote Execution Queue Window

This window contains the following elements:

Status

The following information is available for each job:

- **Job#—**Shows the unique identifier assigned by Signet for each job.
- **Job Name—**Shows the name of the script sent to the server.
- **Genome—**Shows the genome from which the data to be analyzed originates.
- **Time Submitted—**Shows the time that the user launched the script from GeneSpring.
- **Time Started—**Shows the time that the remote execution server began executing the script. If the script is still waiting to be executed this column reads “Pending.” The column reads “Suspended” if the script was paused.
- **Time Finished—**Shows the time that the execution server finished running the script.
- **Actual Running Time**—Shows the actual running time of the script.
- **Estimated Wait Time**—Shows the estimated wait time until the script is run.
- **Estimated Run Time**—Shows the estimated time to run the script.
- **Notify me when results are available**—Specifies whether to display a window informing you when results are available.

Buttons

- **Pause—**Pauses the execution of a pending script so that another script can run first. You cannot pause a script once it has begun executing.
- **Resume—**Resumes the paused script.
- **Refresh—**Refreshes the status of the queue. This window does not refresh automatically. You need to click the Refresh button to see the most recent status of your scripts.
- **View Script—**Displays the selected script.
- **View Results—**Retrieves the results of a script after it has finished running on the remove server. Click this button to retrieve and save the results of your script. A series of file-dialog windows appears that correspond to each of the outputs that your script generates.
- **Close**—Closes the window.

Running Scripts Remotely

To run a script remotely:

1. Open the **Run Script** window.

   Go to "Opening the Run Script Window" on page 16-6 for instructions.

2. Select the **Compute on a Signet Server** option.
3. Click the **Start** button.

   This button is not active until all required data has been entered. Your script is either sent to an available remote server or it is placed in a queue.

   - If your script is supposed to upload data to a regulatory compliant Signet, you will be prompted to provide an electronic signature.
   - If your script produces results that generate a Name Problems window, go to "Handling Scripts that Generate Long Names or Invalid Characters" on page 16-12 for more information.

4. If one or more Results windows open and show a data object, do one of the following:

   - To save your results, enter a **Name**, add any comments in the **Notes** field, and then click the **Save** button.
   - If you do not want to save the results, click the **Cancel** button.

Viewing the Results of a Remote Job

To view the results of a remote job:

1. Select **Tools > Check Remote Execution Queue**.

   The Remote Execution window opens.

2. Select the script you want to view and click the **View Results** button.
3. Do one of the following:
   - To save your results, enter a **Name**, add any comments in the **Notes** field, and then click the **Save** button.
   - If you do not want to save the results, click the **Cancel** button.
4. Click **Close** when you are done.

Handling Scripts that Generate Long Names or Invalid Characters

If your script generated names longer than 80 characters or invalid characters, the Name Problems window opens (Figure 16-5).

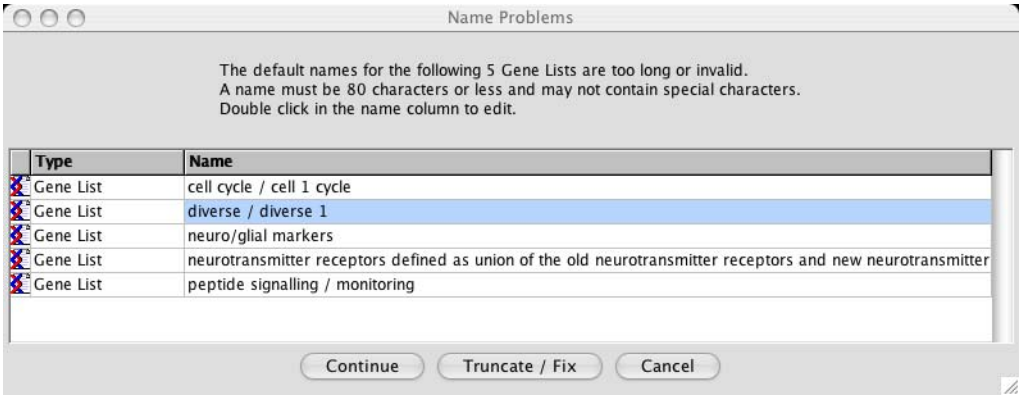

Name Problems
The default names for the following 5 Gene Lists are too long or invalid.
A name must be 80 characters or less and may not contain special characters.
Double click in the name column to edit.
Type
Name
Gene List
cell cycle / cell 1 cycle
Gene List
diverse / diverse 1
Gene List
neuro/glial markers
Gene List
neurotransmitter receptors defined as union of the old neurotransmitter receptors and new neurotransmitter
Gene List
peptide signalling / monitoring
Continue
Truncate / Fix
Cancel

Figure 16-5 Name Problems Window

This window contains the following elements:

- **Table**—Displays the results that have name problems.
- **Continue**—Disabled until the problems are fixed.
- **Truncate/Fix**—Truncates or fixes all of the problem entries based on their allowable size and characters.
- **Cancel**—Cancels the operation and closes the window.

To fix long names or invalid characters:

1. Do one of the following:
   - Manually fix all of the problem entries by editing them.
   - Automatically fix all of the problem entries by clicking the **Truncate/Fix** button.
2. Click the **Continue** button to fix the problem.

Uploading Scripts to Signet

To upload a script to Signet:

- Do one of the following:
  - In the **Navigator**, open the **Scripts** folder, right-click the script you want, and then select the **Upload to Signet** command.
  - In the **Navigator**, open the **Scripts** folder, drag the script you want, and drop it on the Signet folder you want.

Inspecting Scripts

This section explains how to use the Script Inspector to inspect and edit script details.

Script Inspector Window

The Script Inspector window (Figure 16-6) lets you examine the details of a particular script. The contents of this window will vary depending on the type of script you are inspecting.

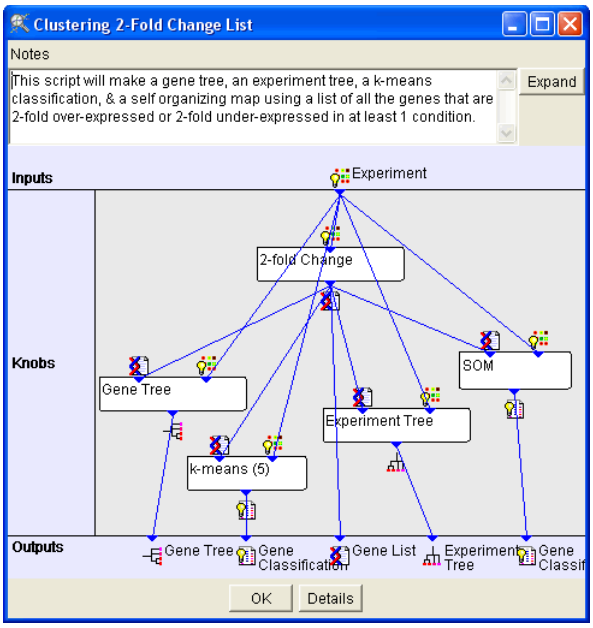


Figure 16-6 Script Inspector Window

The Script Inspector window contains the following elements.

Notes

The Notes field displays notes added to the script. To enter or edit notes, you need to use the ScriptEditor.

Buttons

- **OK**—Closes the window.
- **Details**—Opens a new window that shows details about the script.

Opening the Script Inspector Window

To open the Script Inspector window:

- Do one of the following:
  - In the **Navigator**, open the **Scripts** folder, right-click the script you want, and then select the **Inspect** command.
  - Open the Run Script window and then click the **View Script** button.

    Go to "Opening the Run Script Window" on page 16-6 for instructions.

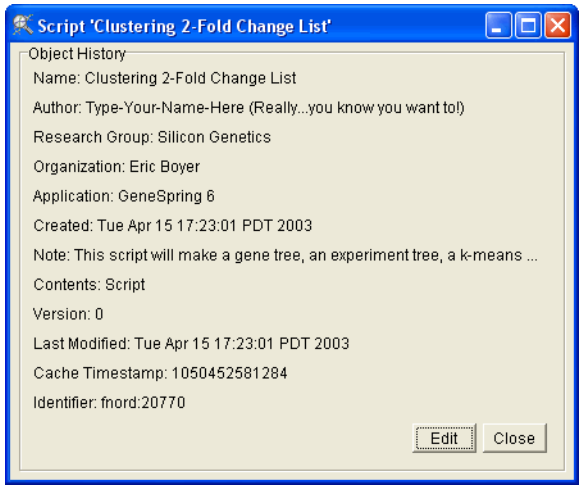
Editing Script Details

To edit script details:

1. Open the **Script Inspector**.

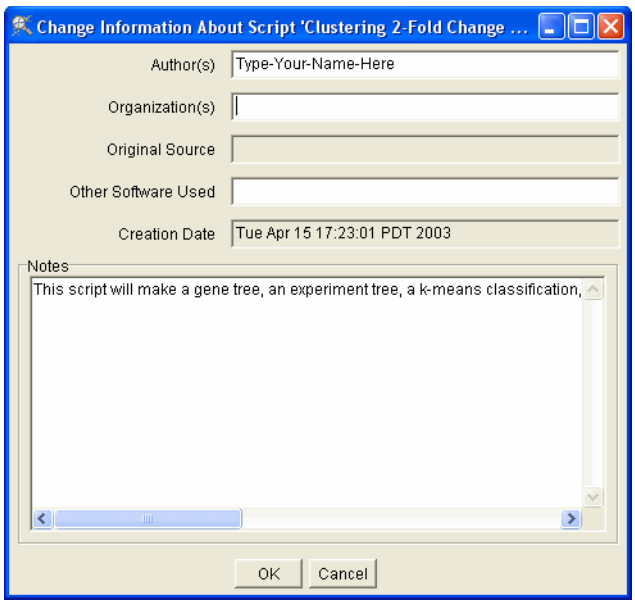
   Go to "Opening the Script Inspector Window" on page 16-14 for instructions.
2. Click the **Details** button.

   The Details window opens.


Script 'Clustering 2-Fold Change List'
Object History
Name: Clustering 2-Fold Change List
Author: Type-Your-Name-Here (Really...you know you want to!)
Research Group: Silicon Genetics
Organization: Eric Boyer
Application: GeneSpring 6
Created: Tue Apr 15 17:23:01 PDT 2003
Note: This script will make a gene tree, an experiment tree, a k-means ...
Contents: Script
Version: 0
Last Modified: Tue Apr 15 17:23:01 PDT 2003
Cache Timestamp: 1050452581284
Identifier: fnord:20770
Edit
Close

3. Click the **Edit** button.

The Change Details window opens.



4. Change the **Authors**, **Organization**, **Original Source**, **Original Software Used**, or **Notes** fields.
5. Click the **OK** button.

Using the Script Editor

The Script Editor lets you create and modify your own scripts in GeneSpring. All scripts, including complimentary scripts shipped with GeneSpring, are stored in the Scripts folder in the Navigator. This section covers the following topics:

- Script Editor Terminology
- Script Editor Window
- Opening the Script Editor Window

Script Editor Terminology

Before using the Script Editor, you might want to review the basic terminology associated with a script.

- **Building block**—One of the building blocks used to create scripts; for example, a script primitive, script, or external program building block.
- **Script**—A more complex program made up of script building blocks, other scripts, and/or external program building blocks.
- **External program building block**—A building block that is created for each external program defined within GeneSpring. A script with an external program building block can only be run on a version of GeneSpring that has the external program installed. Or,

in certain cases, an external program building block can be run on a remote server if the building block does not require a graphical user interface (because it cannot be shown on a remote server).

For more information on GeneSpring's External Program Interface, see "Running External Programs" on page 16-56.

- **Socket**—Parts of a building block that send or receive a connection from another building block. Inputs and outputs are sockets.
- **Building block input**—Parts of a building block that receive information. Inputs are located on the top of a building block, and receive connections only from building block outputs or script inputs.
- **Building block output**—Parts of a building block that send the results of the operation performed. Outputs are located on the bottom of a building block, and receive connections only from building block inputs or script outputs.
- **Script input**—The socket at the top of a script. Script inputs are created when you connect a line from an input socket to the "Inputs" area of the Browser.
- **Script output**—The final output of a script. Script outputs are created when you drag a building block output into the "Outputs" area of the browser.
- **Knob**—A placeholder for a value entered when running the script.

Script Editor Window

The Script Editor window (Figure 16-7) lets you automate the routine tasks of analyzing, interpreting, and archiving large volumes of expression data. By dragging and dropping icons to a canvas, you can record the steps of complex analyses in a simple script or combine existing scripts to create a more powerful script.

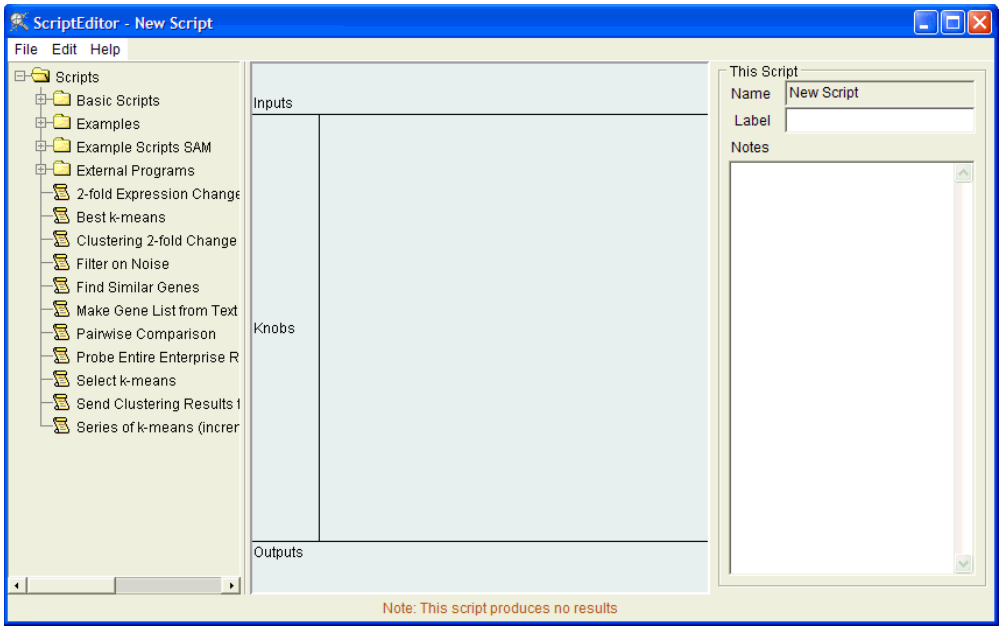


Figure 16-7 Script Editor Window

You can also share or exchange scripts with colleagues, maintaining consistency in the analysis process and simplifying data management. Computationally-intensive analyses can easily be captured and off-loaded to Signet's remote servers, via GeneSpring's Remote Execution feature.

The Script Editor window contains a Navigator panel, a Browser panel, and a panel that switches dynamically between Notes and Blocks. This section describes the elements in these panels.

Navigator Panel

The Navigator panel contains the building blocks, scripts, or external program building blocks you use to create your scripts. On Windows, click the + (plus sign) next to a folder to open it, or click the - (minus sign) to close it. On Macintosh, click the triangle to open or close a folder.

Browser Panel

The Browser (Figure 16-8) lets you create and browse scripts.

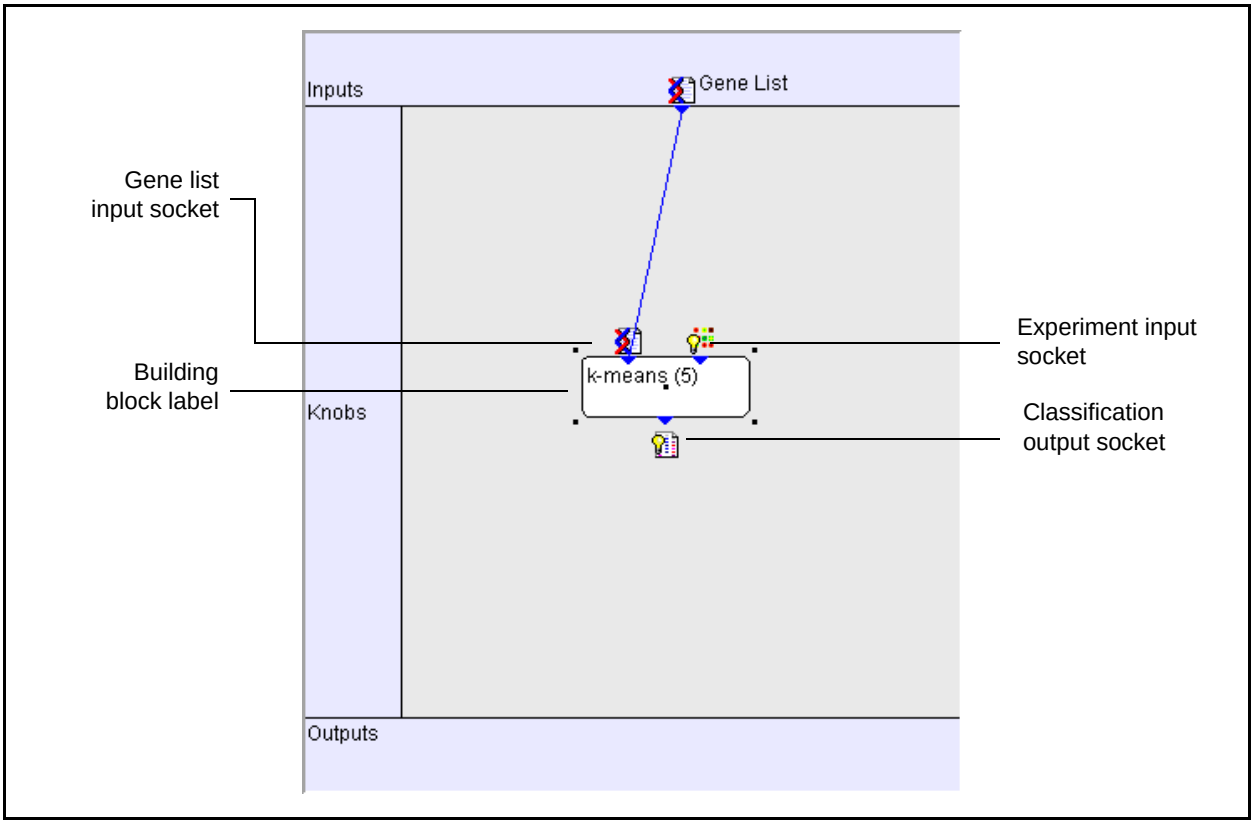


Figure 16-8 Script Editor Window: Browser Panel

You can select any line in the panel and drag it to move it. Moreover, long labels can be displayed by hovering the mouse over the label to reveal the text.

Notes Panel

The Notes panel, located on the right side of the Script Editor, lets you add or edit notes to the script.

This Script Section

The **This Script** section displays the following information about the currently selected item:

- **Name**—The name of the selected script.
- **Label**—A label for the script, distinct from the script name. You can enter a brief description of the script to make it easier to identify. The Label is displayed instead of the Script Name when the script is used as a block inside another script.
- **Notes**—You can alter these notes in any way you wish and add a nearly unlimited amount of text. These notes will be associated with the script.

If no item is selected, empty boxes are displayed.

Block Panel

Selecting a block in the Browser displays information about the building block (Figure 16-9) in the right corner of the Script Editor. It replaces the Notes panel.

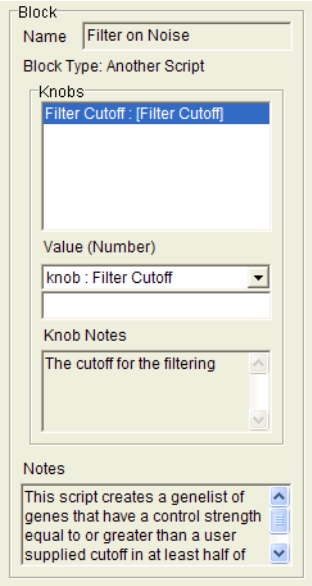

Block
Name
Filter on Noise
Block Type: Another Script
Knobs
Filter Cutoff : [Filter Cutoff]
Value (Number)
knob : Filter Cutoff
Knob Notes
The cutoff for the filtering
Notes
This script creates a genelist of genes that have a control strength equal to or greater than a user supplied cutoff in at least half of

Figure 16-9 Script Editor Window: Block Panel

In Figure 16-9, the selected item in the browser is the primitive “Filter Genes with Associated Numbers”. This figure shows information about the knobs associated with the selected building block. These items are different for each building block. You might see several types of input fields including menus and text boxes.

The Block section contains the following information:

- **Name**—The name of the selected building block
- **Misc**—This box can display a variety of information about the selected building block. It only appears for certain building blocks.
- **Notes**—The function of the selected building block.

Help

You can get context-sensitive help on any script building block by right-clicking on it and selecting **Help**. Double-clicking a building block brings up an inspector window for that building block.

Icon Legend

There are many icons in the Script Editor intended to help you keep track of the different objects that are available (Figure 16-10). To view information on these icons, select **Help > Icon Legend**. You may want to leave the Icon Legend open and visible on your desktop until you are accustomed to working with the icons.

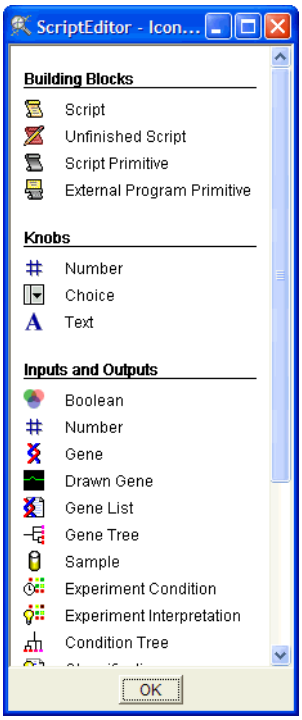


Figure 16-10 Script Editor Window: Icon Legend

Script Inspector

The Script Inspector lets you view properties associated with scripts, building blocks, history, and change information. For example, right-clicking on a script in the Navigator and selecting the Inspect command displays properties for that script.

Go to "Script Inspector Window" on page 16-13 for more information.

Opening the Script Editor Window

To open the Script Editor window:

- Do one of the following:
  - In the **Navigator**, open the **Scripts** folder, right-click the script you want, and then select the **Edit** command.
  - Select **Tools** > **Script Editor**.

Working with Building Blocks

Scripts can be created out of the most basic scripting elements, known as building blocks, as well as out of other scripts. You can combine many simple scripts to create more complex scripts. This section includes the following topics:

- Building Block Types
- Inputs and Outputs
- Knobs
- Sample Script
- Pre-defined Building Blocks
- Opening Building Blocks
- Opening Pre-defined Building Blocks
- Using Building Blocks from External Programs

Building Block Types

Four types of data objects can be used as building blocks in the creation of scripts:

- Scripts
- Script building blocks
- External program building blocks
- Script plugins

Within a script, data is passed from one building block to another as the script runs. To create a new script, drag one or more building blocks from the Navigator to the browser, then drag the cursor from the input socket of one building block to the matching output socket of another building block, or vice versa. A line appears between the two sockets.

GeneSpring also provides a set of pre-defined building blocks for you to use. Go to "Pre-defined Building Blocks" on page 16-23 for more information.

Inputs and Outputs

There are two types of inputs and outputs: building blocks inputs and outputs and script inputs and outputs.

Building Block Inputs and Outputs

Building block inputs and outputs specify data flow through a script.

Script Inputs and Outputs

Script inputs are created when you drag a building block input into the “Inputs” area of the browser. Script outputs are created when you drag a building block output into the “Outputs” area of the browser.

There are three types of information produced by scripts that cannot be saved in GeneSpring. This is because there is no way to manipulate this information within GeneSpring.

- **Boolean**—A simple results box appears for this output, typically either true or false. See “Boolean” on page 16-24 for details.
- **Numbers**—A simple results box appears containing a list of numbers for this output. See “Numbers” on page 16-30 for details.
- **Sequence Information**—This information is displayed in a copy-and-paste enabled table, similar to the Potential Regulatory Sequences table in GeneSpring.

Knobs

Knobs allow the user running the script to enter a value when the script is run. For example, you could filter with a knob that sets the minimum normalized expression level.

Information for inputs, outputs, and knobs is entered in GeneSpring at the time you run the script. Knobs required for building blocks are specified in the Script Editor Block area on the right side of the Script Editor window.

Sample Script

Scripts are very simple when stripped of their fancy verbiage and icons. Figure 16-11 shows an example of a script flow chart to examine your data.

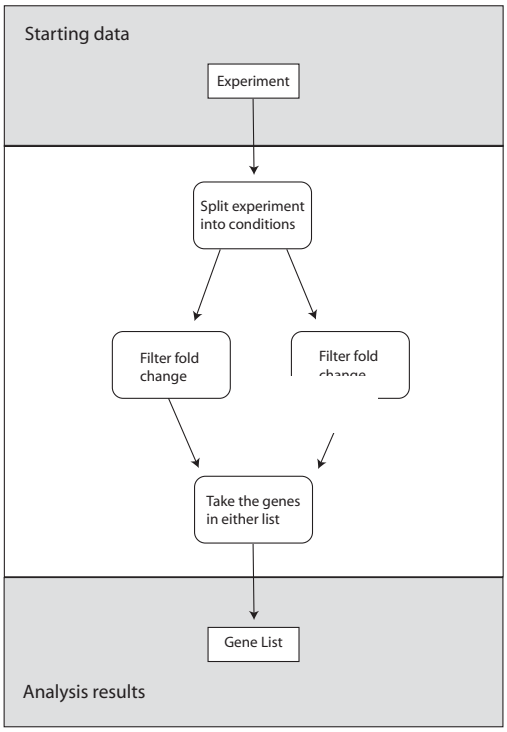


Figure 16-11 Script Flow Chart

Figure 16-12 shows what the example script flow chart would look like in the Script Editor.

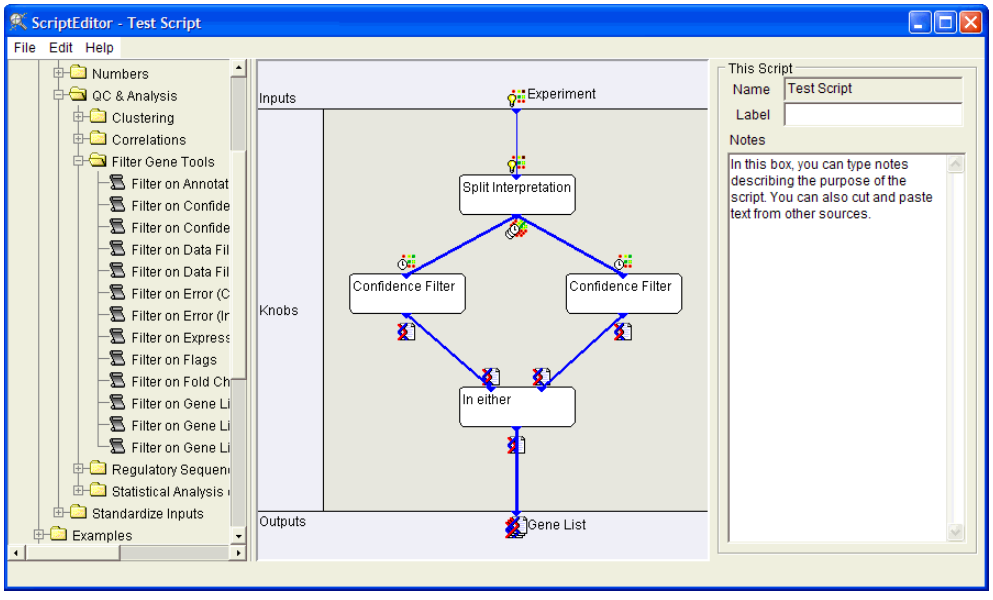


Figure 16-12 Script Flow Chart in Script Editor

Everything in the browser must be correctly connected to another object. Check the bottom of the window for error or warning messages. If any error messages are present, the script cannot be run. You can run a script with warning messages active, but it may not function as intended.

Pre-defined Building Blocks

The Script Editor comes with a predefined set of building blocks (also referred to as script primitives) you can join together in various ways to build scripts. There are several categories of building blocks:

- Boolean
- Select using Boolean
- Gene List Manipulations (Venn Diagram)
- Signet Downloading (Default Directory)
- Signet Downloading (Specified Directory)
- Signet Publishing (Default Directory)
- Signet Publishing (Specific Directory)
- Look Up
- Merge-Split Groups
- Make Groups
- Select Groups
- Numbers
- Count Groups
- Clustering
- Correlations
- Filter Gene Tools
- Regulatory Sequence Search
- Statistical Analysis (ANOVA)
- Standardize Inputs
- External Programs

You can combine various building blocks to create a script. For very long or complex scripts, you may want to create several small scripts and join them together in the Script Editor to create the final script.

This section describes the pre-defined building blocks that are available to use in GeneSpring.

Boolean

Table 16-2 describes the Boolean building blocks that are available.

Table 16-2 Boolean Building Blocks

Name	Input	Output	Description
Boolean	No direct input	Boolean	Generates a *true* or *false* result. Select *yes* (true) or *no* (false) when you run the script.
Boolean AND	At least two Booleans	Boolean	Outputs is *true* only if both inputs are *true*.
Boolean FALSE	None	Boolean	Returns the result *false*.
Boolean NOT	One Boolean	Boolean	Outputs is *true* only if the input is *false* (converts true to false and false to true).
Boolean OR	Two Booleans	Boolean	Outputs is *true* if either input is *true*.
Boolean TRUE	None	Boolean	Returns the result *true*.

Select using Boolean

Table 16-3 describes the Select using Boolean building blocks that are available.

Table 16-3 Select using Boolean Building Blocks

Name	Input	Output	Description
Select Boolean	Three Booleans	Boolean	Selects the second boolean input if the first input is true and selects the third boolean input if the first input is false.
Select Condition	One Boolean, two conditions	Condition	Selects the first condition if true, the second condition if false.
Select Condition Tree	One Boolean, two condition trees	Condition tree	Selects the first tree if true, the second tree if false.
Select Experiment	One Boolean, two experiment interpretations	Interpretation	Selects the first interpretation if true, the second if false.
Select Gene	One Boolean, two genes	Gene	Selects the first gene if true, the second if false.
Select Gene Classification	One Boolean, two classifications	Classification	Selects the first classification if true, the second if false.
Select Gene List	One Boolean, two gene lists	Gene list	Selects the first gene list if true, the second if false.
Select Gene Tree	One Boolean, two gene trees	Gene tree	Selects the first tree if true, the second if false.
Select Number	One Boolean, two numbers	Number	Selects the first number if true, the second if false.
Select Sample	One Boolean, two samples	Sample	Selects input 1 if the Boolean is true, and 2 if the Boolean is false.

Table 16-3 Select using Boolean Building Blocks (Continued)

Name	Input	Output	Description
Select Sequence	One Boolean, two sequences	Sequence	Selects the first sequence if true, the second if false.

Gene List Manipulations (Venn Diagram)

Table 16-4 describes the Gene List Manipulations (Venn Diagram) building blocks that are available.

Table 16-4 Gene List Manipulations (Venn Diagram) Building Blocks

Name	Input	Knobs	Description
All Genes	None	None	Outputs the list of all genes.
All Genomic	None	None	Outputs a list of all genomic elements.
Count Genes in Gene List	At least one gene list	None	Outputs the number of genes in the input lists.
Gene List Difference	Two gene lists	None	Outputs a list of the genes that are in the first gene list, but not the second.
Gene List Intersection	Two gene lists	None	Outputs a list of the genes that are in both input lists.
Gene List Inversion	Two gene lists	None	Outputs a gene list that contains all of the genes in the genome, except genes that are in the input gene list.
Gene List Union	Two gene lists	None	Outputs a list of the genes that are in either input list.
In All Gene Lists	One gene list group	None	Outputs a list of the genes in all of the input lists.
In at Least One	One gene list group	None	Outputs a list of the genes in at least one of the input lists.
In a Number of Gene Lists	Array of gene lists	Number of gene lists from the array a gene must be in, Comparison (>, <, =, <=, >=)	Outputs a list of genes that appear in at least (or other comparison) the specified number of gene lists.
In a Percentage of Gene Lists	Array of gene lists	Percentage, Comparison	Outputs a list of genes that appear in a specified proportion of the gene lists.
Sort Gene List	One gene list	none	Outputs the gene list sorted in descending order based on its associated numbers.

Signet Downloading (Default Directory)

Table 16-5 describes the Signet Downloading (Default Directory) building blocks that are available. You'll need to login to Signet before using a script to download data from Signet.

Table 16-5 Signet Downloading (Default Directory) Building Blocks

Name	Inputs	Knobs	Description
Download a Gene List from Signet	None	Gene List name	Outputs a specified gene list retrieved from Signet.
Download an Experiment from Signet	None	Experiment name	Outputs a specific experiment retrieved from Signet.
Download all Gene Lists from Signet	None	None	Outputs all gene lists retrieved from Signet.
Download All Experiments from Signet	None	None	Outputs all experiment interpretations retrieved from Signet.

Signet Downloading (Specified Directory)

Table 16-6 describes the Signet Downloading (Specified Directory) building blocks that are available. You'll need to login to Signet before using a script to download data from Signet.

Table 16-6 Signet Downloading (Specified Directory) Building Blocks

Name	Inputs	Knobs	Description
Download All Gene Lists from Directory in Signet	None	Directory name	Outputs an array of gene lists retrieved from the specified folder in Signet.
Download All Experiments from Directory in Signet	None	Directory name	Outputs an experiment or array of experiments retrieved from the specified folder in Signet. This lets you hard-code a reference to an experiment or folder of experiments in your script.

Signet Publishing (Default Directory)

Table 16-7 describes the Signet Publishing (Default Directory) building blocks that are available. You'll need to login to Signet before using a script to autoload data to Signet.

Table 16-7 Signet Publishing (Default Directory) Building Blocks

Name	Input	Knobs	Description
Send Classification to Signet	One classification	None	Publishes a classification to your default directory in Signet. No output to GeneSpring.
Send Experiment to Signet	one interpretation	None	Publishes an experiment interpretation to your default directory in Signet. No output to GeneSpring.
Send Condition Tree to Signet	One condition tree	None	Publishes an condition tree to your default directory in Signet. No output to GeneSpring.
Send Gene List to Signet	One gene list	None	Publishes a gene list to your default directory in Signet. No output to GeneSpring.
Send Gene Tree to Signet	One gene tree	None	Publishes a gene tree to your default directory in Signet. No output to GeneSpring.

Signet Publishing (Specific Directory)

Table 16-7 describes the Signet Publishing (Specified Directory) building blocks that are available.You'll need to login to Signet before using a script to autoload data to Signet

Table 16-8 Signet Publishing (Specified Directory) Building Blocks

Name	Input	Knobs	Description
Send Classification to Directory in Signet	One classification	Directory	Publishes a classification to a chosen directory in Signet. No output to GeneSpring.
Send Experiment to Directory in Signet	one interpretation	Directory	Publishes an experiment interpretation to a chosen directory in Signet. No output to GeneSpring.
Send Condition Tree to Directory in Signet	One condition tree	Directory	Publishes an condition tree to a chosen directory in Signet. No output to GeneSpring.
Send Gene List to Directory in Signet	One gene list	Directory	Publishes a gene list to a chosen directory in Signet. No output to GeneSpring.
Send Gene Tree to Directory in Signet	One gene tree	Directory	Publishes a gene tree to a chosen directory in Signet. No output to GeneSpring.

Look Up

Table 16-9 describes the Look Up building blocks that are available.

Table 16-9 Look Up Building Blocks

Name	Input	Description
Is Gene in Gene List	One gene, one gene list	Returns true if the gene list contains the input gene.
Number Associated with Gene in Condition	One gene, one condition	Returns the number associated with a gene in a condition. Genes without a number will be ignored.
Number Associated with Gene in Gene List	At least one gene, one gene list	Returns the number associated with a gene in a gene list. Genes without a number will be ignored.
Interpretation Restriction	one interpretation	Takes an experiment interpretation and produces a temporary interpretation consisting of parameters and values specified in the string restriction.

Merge-Split Groups

Table 16-10 describes the Merge-Split Groups building blocks that are available.

Table 16-10 Merge-Split Groups Building Blocks

Name	Input	Description
Merge Genes	One gene group	Outputs a list containing all lists in the group input.
Merge Genes and Numbers	One gene group, one number group	Merges a group of genes into a gene list with associated numbers. The resulting gene list will contain only the genes with matching numbers. Unmatched genes and numbers will not be included in the resulting gene list.
Split Classification	One classification	Splits the classification into a group of gene lists.
Split Condition	One interpretation	Splits the condition up into a group of samples.
Split Gene List	One gene list	Splits the gene list into a collection of individual genes.
Split Gene List With Numbers	At least one gene list	Splits the gene list into a group of genes and an associated group of numbers.
Split Interpretation	One interpretation	Splits the interpretation up into a group of conditions or samples.

Make Groups

Table 16-11describes the Make Groups building blocks that are available**.**

Table 16-11 Make Groups Building Blocks

Name	Input	Description
Make Classification Group	A folder of classifications	Produces a group of classifications.
Make Experiment Group	A folder of experiments	Outputs a group of interpretations. There is one knob, to select whether to include all the interpretations or only the defaults.
Make Gene List Group	A folder of gene lists	Outputs a group of gene lists.
Make Gene Tree Group	A folder of gene trees	Outputs a group of gene trees.

Select Groups

Table 16-12 describes the Select Groups building blocks that are available.

Table 16-12 Select Groups Building Blocks

Name	**Input**	**Description**
Filter Boolean Group	Two Boolean groups	If true, passes the second argument for each Boolean through the corresponding first argument.
Filter Condition Group	One Boolean group and one condition group	If true, passes the second argument for each Boolean through the corresponding first argument.
Filter Experiment Group	One Boolean group, one interpretation	If true, outputs an experiment interpretation for each Boolean in the first argument.
Filter Condition Tree Group	One Boolean group, one condition tree group	If true, outputs an condition tree for each Boolean in the first argument.
Filter Gene Classification Group	One Boolean group, one classification group	If true, outputs a classification for each Boolean in the first argument.
Filter Gene Group	One Boolean group, one gene group	If true, outputs a gene for each Boolean in the first argument.
Filter Gene List Group	One Boolean group, one gene list group	If true, outputs a gene list for each Boolean in the first argument.
Filter Gene Tree Group	One Boolean group, one gene tree group	If true, outputs a gene tree for each Boolean in the first argument.
Filter Number Group	One Boolean group, one number group	If true, outputs a number for each Boolean in the first argument.
Filter Sequence Group	One Boolean group, one sequence group	If true, outputs a sequence for each Boolean in the first argument.

Numbers

Table 16-13 describes the Numbers building blocks that are available.

Table 16-13 Numbers Building Blocks

Name	Input	Knobs	Description
Compare 1 Number	One number	Comparison, Number	Compares a number to another number and outputs a Boolean.
Compare 2 Numbers	Two numbers	Comparison	Compares two numbers and outputs a Boolean.
List of Numbers	None	String of Numbers (this must contain a comma-delimited list of numbers, or number dash number tokens. Decimals cannot be used in conjunction with dashes.)	Outputs a group of numbers from a user specified comma delimited list. For example: l 2,3,5,10-12 will produce 1 2 3 5 10 11 12.
Number	None	Number	Outputs the number specified by a knob.
Number Add	Two numbers	None	Adds two numbers together and outputs the result.
Number Divide	Two numbers	None	Divides the first number by the second number and outputs the result.
Number Multiply	Two numbers	None	Multiplies two numbers and outputs the result.
Number Subtract	Two numbers	None	Subtracts the second number from the first number and outputs the result.
Range of Numbers	Starting Number, Step Value, How Many Numbers	None	Outputs a list of numbers that form a linear progression.
Number Log	Number	None	Outputs the base 10 log of the number specified in the input.
Sum of Numbers in Group	A group of numbers	None	Outputs the sum of the numbers in the input group.

Count Groups

Table 16-14 describes the Count Groups building blocks that are available.

Table 16-14 Count Groups Building Blocks

Name	Input	Knobs	Description
Count Conditions in Group	Array of Conditions	None	Determines the number of objects in the specified array and outputs the result.
Count Experiments in Group	Array of Experiments	None	Determines the number of objects in the specified array and outputs the result.
Count Gene Lists in Group	Array of Gene Lists	None	Determines the number of objects in the specified array and outputs the result.
Count Sequences in Group	Array of sequences	None	Determines the number of objects in the specified array and outputs the result.

Clustering

Table 16-15 describes the Clustering building blocks that are available.

Table 16-15 Clustering Building Blocks

Name	Input	Knobs	Description
Build Condition Tree	At least one gene list, one interpretation	Correlation type, Separation ratio, Minimum distance	Outputs a condition tree.
Build Gene Tree	One gene list, one interpretation	Similarity measure, Merge similar branches, Discard bad, Separation ratio, Minimum distance, Do automatic annotation, Use standard	Outputs a gene tree.
Explained Variation	At least one classification, one interpretation, one gene list.	None	Computes the proportion of variation in an experiment interpretation explained by a classification and a gene list. Output is a number between zero and one inclusive (for example, 0.14567 is 14.567% explained variability).
Find Predictor Genes	one interpretation, one gene list	Parameter name, number of genes	Outputs a list of genes that are good at predicting a given parameter in an experiment.

Table 16-15 Clustering Building Blocks (Continued)

Name	Input	Knobs	Description
K-means	One gene list and one interpretation	Number of groups, Similarity measure, Maximum iterations, Additional tries, and Discard bad.	Outputs a k-means classification.
K-means with Starting Classification	One gene list, one interpretation, one, one classification	Similarity measure, Number of iterations, and Discard bad.	Outputs a k-means clustering starting from an existing classification.
Self-Organizing Map	One gene list, one interpretation	Iterations, Discard bad, Rows, Columns, and Radius.	Outputs a self-organizing map (SOM).
Principal Component Analysis (Conditions)	One gene list, one condition	Report scores as correlations	Identifies predominant expression patterns in the input condition.
Principal Component Analysis (Genes)	One gene list, one experiment	Report scores as correlations	Identifies predominant expression patterns in the input experiment.
QT Clustering	One gene list, one experiment	Minimum cluster size, minimum correlation, and a similarity measure	Outputs QT clustering.

Correlations

Table 16-16 describes the Correlations building blocks that are available.

Table 16-16 Correlations Building Blocks

Name	Input	Knobs	Description
Condition Correlation	Two conditions, one gene list	Correlation	Compares two conditions looking at only those genes in the specified gene list, and outputs a p-value.
Find Similar Conditions in Experiment	One condition, one experiment, one gene list	Correlation, Cut-off value	Compares the specified condition to every condition in the specified experiment, using only the genes in the specified gene list. Outputs an array of conditions and an associated array of p-values.

Table 16-16 Correlations Building Blocks (Continued)

Name	Input	Knobs	Description
Find Similar Genes	One gene, one gene list, one or more experiment interpretations, one or more weights	Similarity measure, Minimum correlation, Maximum correlation	Outputs a list of genes whose expression profiles are correlated to a specified gene over the conditions of an interpretation.
Find Similar Genes (with custom input)	One gene, one gene list, one or more experiment interpretations, one or more weights	Similarity measure, Minimum correlation, Maximum correlation, Conditions usage	Outputs a list of genes whose expression profiles are correlated to a specified gene over the conditions of an interpretation.
Find Similar Samples	One gene list, one target sample, one sample pool	Apply weights, Weighting coefficient	Outputs a list of samples correlated to a specified sample.
Gene Correlation	Two genes, one experiment	Correlation	Compares the two genes determine their similarity with respect to the selected experiment. Output a p-value.
Gene List Similarity p-Value	Two gene lists	None	Calculates the similarity between two gene lists, using the All Genes list as the Universe. Outputs a number representing the probability that the intersection between the two lists could be due to chance. Note that there is no multiple testing correction applied by this script.
Gene List Similarity p-Value, Specified Universe	Three gene lists	None	Calculates the similarity between two gene lists, using the specified third gene list as the Universe. Outputs a number representing the probability that the intersection between the two lists could be due to chance. Note that there is no multiple testing correction applied by this script.
Find Significant Parameters (ANOVA)	One gene list, one experiment	P-value cutoff, test type	Finds parameters and attributes correlated to expression using an ANOVA.
Find Significant Parameters (Association Test)	One gene list, one experiment	P-value cutoff	Finds parameters and attributes correlated to expression using an association test.
Find Significant Parameters (Correlation to Parameters)	One gene list, one experiment	P-value cutoff, enumeration	Finds numeric parameters and attributes that are correlated to expression using a simple correlation.

Filter Gene Tools

Table 16-17 describes the Filter Gene Tools building blocks that are available.

Table 16-17 Filter Gene Tools Building Blocks

Name	Input	Knobs	Description
Filter on Annotations	Gene list	text string, other knobs in script	Outputs a list of genes whose annotations contain a specified text string.
Filter on Confidence (Condition Input)	One condition, one gene list	Measure of confidence, Multiple testing correction, Minimum, Maximum	Filters on t-test p-value or number of replicates using a condition. Outputs a gene list.
Filter on Confidence (Interpretation Input)	One interpretation, one gene list	Measure of confidence, Multiple testing correction, Minimum, Maximum, Minimum # of conditions	Filters on t-test p-value or number of replicates using an experiment interpretation. Outputs a gene list.
Filter on Data File (Condition Input)	One condition	Filter method, Filter text, Use * as wildcard, Must appear in, # of samples, Filter columns specification	Filters on a data file using a condition. Outputs a gene list.
Filter on Data File (Interpretation Input)	One interpretation	Filter method, Filter text, Use * as wildcard, Must appear in, # of samples, Filter columns specification	Filters on a data file using an experiment interpretation. Outputs a gene list.
Filter on Error (Condition Input)	One condition	Error type, Minimum, Maximum	Filters on errors using a condition. Outputs a gene list.
Filter on Error (Interpretation Input)	One interpretation	Error type, Minimum, Maximum, Minimum # of conditions	Filters on errors using an experiment interpretation. Outputs a gene list.
Filter on Expression Level (Array of Condition)	array of conditions	Data type, Minimum, Maximum, Minimum # of conditions	Filters on expression level using an array of conditions. Outputs a gene list.

Table 16-17 Filter Gene Tools Building Blocks (Continued)

Name	Input	Knobs	Description
Filter on Expression Level (Single Condition)	One condition input	Data type, Minimum, Maximum	Outputs a gene list containing the genes that have a measurement relative to a cutoff.
Filter on Flags	One or more samples	Flag value, Minimum # of samples	Filters on flags and outputs a gene list.
Filter on Fold Change	One condition, one condition group	Data type, Comparison, Fold difference, Must appear in	Filters on fold change and outputs a gene list.
Filter on Gene List	Gene list	None	Outputs the input gene list.
Filter on Gene List Numbers	Gene list with associated numbers	Cutoff, Comparison	Produces a gene list (from an existing gene list) containing the genes whose associated number meets the specified criteria.
Filter on Gene List Numbers (In Range)	Gene list with associated numbers	Upper bound, Lower bound	Produces a gene list (from an existing gene list) containing genes whose associated number meets the specified criteria.
Filter on Parameter	One experiment, one gene list	Min, Max, Parameter Name, and Similarity Measure	Computes a correlation between an experimental parameter and expression level data.
Filter on Sample Attribute	One experiment, one gene list	Min, Max, Attribute Name, Attribute Units, and Similarity Measure.	Computes a correlation between a sample attribute and expression level data.

Regulatory Sequence Search

Table 16-18 describes the Regulatory Sequence Search building blocks that are available.

Table 16-18 Regulatory Sequence Search Building Blocks

Name	Input	Knobs	Description
Find Genes with Specific Regulatory Sequence	One sequence, one gene list	From Base, To Base, Maximum errors, Relative Genomic, Local Correction	Outputs a list of the genes that contain the input regulatory sequence.
Find Genes with Specific Regulatory Sequence in Universe	One gene list, one sequence, and a Universe of genes	From Base, To Base, Maximum errors, Relative Genomic, and Local Correction	Outputs a list of the genes that contain the input regulatory sequence.

Table 16-18 Regulatory Sequence Search Building Blocks (Continued)

Name	Input	Knobs	Description
Find Regulatory Sequences	One gene list	From Base, To Base, Minimum Length, Maximum Length, Maximum Errors, Minimum Interior N's, Relative Genomic, P-value Cutoff, Local Correction	Outputs a list of regulatory sequences upstream of the genes in the input list.
Find Regulatory Sequences in Universe	One gene list, Universe of genes	From Base, To Base, Minimum Length, Maximum Length, Maximum Errors, Minimum Interior N's, Relative Genomic, P-value Cutoff, Local Correction	Find regulatory sequences upstream of the genes specified as input.

Statistical Analysis (ANOVA)

Table 16-19 describes the Statistical Analysis (ANOVA) building blocks that are available.

Table 16-19 Statistical Analysis Building Blocks

Name	Input	Knobs	Description
1-way ANOVA	One gene list, one interpretation	Groups specification, Test type, P-value cutoff, Multiple testing correction	See "Performing One-way ANOVA" on page 13-5 for details.
1-way ANOVA with Post Hoc Tests	One gene list, one interpretation	Groups specification, Test type, P-value cutoff, Multiple testing correction, Post hoc tests	See "Performing One-way ANOVA" on page 13-5 for details.
2-way ANOVA	One gene list, one interpretation	Groups specification (1st parameter), Groups specification (2nd parameter), Test type, P-value cutoff, Multiple testing correction	See "Performing Two-way ANOVA" on page 13-15 for details.
2-way ANOVA (Specific Result)	One gene list, one interpretation	Groups specification (1st parameter), Groups specification (2nd parameter), Test type, P-value cutoff, Multiple testing correction, Restriction	See "Performing Two-way ANOVA" on page 13-15 for details.

Standardize Inputs

describes the Standardize Inputs building blocks that are available.

Table 16-20 Standardize Inputs Building Blocks

Name	Input	Knobs	Description
Use Specific Experiment	None	Experiment name	Specifies a single local experiment by name. This is useful as an input to another block.
Use Specific Experiment Folder	None	Include subfolders, Experiment folder name	Specifies a single local experiment folder by name. This is useful as an input to another block.
Use Specific Gene List	None	Gene list name	Specifies a single local gene list by name. This is useful as an input to another block.
Use Specific Gene List Folder	None	Include subfolders, Gene list folder name	Specifies a single local gene list folder by name. This is useful as an input to another block.

External Programs

Table 16-21 describes the External Programs building blocks that are available from Silicon Genetics.

Table 16-21 External Programs Building Blocks

Name	Input	Description
Load Classification from File	None	Runs an external program to load and outputs a classification from a file on disk.
Load Experiment from File	None	Runs an external program to load and outputs an experiment from a file on disk.
Load Gene List from File	None	Runs an external program to load and outputs a gene list from a file on disk.
Load Gene List with Numbers from File	None	Runs an external program to load and outputs a gene list with associated numbers from a file on disk.
Load Gene Tree from File	None	Runs an external program to load and outputs a gene tree from a file on disk.
Save Classification to File	One classification	Runs an external program to save a classification to disk. A save data window opens and prompts you to enter a filename.
Save Experiment to File	One experiment	Runs an external program to save a classification to disk. A save data window opens and prompts you to enter a filename.
Save Gene List to File	One gene list	Runs an external program to save a gene list to disk. A save data window opens and prompts you to enter a filename.

Table 16-21 External Programs Building Blocks (Continued)

Name	Input	Description
Save Gene List with Numbers to File	One gene list	Runs an external program to save a gene list with associated numbers to disk. A save data window opens and prompts you to enter a filename.
Save Gene Tree to File	One gene tree	Runs an external program to save a gene tree to disk. A save data window opens and prompts you to enter a filename.

Opening Building Blocks

To open a building block:

1. Select **Tools > Script Editor**.
2. In the **Navigator**, locate the building block you want.
3. Right-click on a building block.
4. Select **Inspect** from the menu.
5. Click **Details** for information on the script's author and creation date.

Opening Pre-defined Building Blocks

Table 16-22 explains how to open building blocks in the Navigator panel of the Script Editor.

Table 16-22 Pre-defined Building Blocks

To open . . .	**Select . . .**
Boolean	Scripts > Basic Scripts > Boolean
Select Using Boolean	Scripts > Basic Scripts > Boolean > Select Using Boolean
Gene List Manipulations	Scripts > Basic Scripts > Gene List Manipulations (Venn Diagram)
Signet Downloading > Specified Directory	Scripts > Basic Scripts > Signet > Signet Downloading > Default Directory
Signet Downloading > Specified Directory	Scripts > Basic Scripts > Signet > Signet Downloading > Specified Directory
Signet Publishing > Default Directory	Scripts > Basic Scripts > Signet Publishing > Default Directory
Signet Publishing > Specified Directory	Scripts > Basic Scripts > Signet Publishing > Specified Directory
Look Up	Scripts > Basic Scripts > Look Up
Merge-Split Groups	Scripts > Basic Scripts > Merge-Split Groups
Make Groups	Scripts > Basic Scripts > Merge-Split Groups > Make Groups
Select Groups	Scripts > Basic Scripts > Merge-Split Groups > Select Groups
Numbers	Scripts > Basic Scripts > Numbers

Table 16-22 Pre-defined Building Blocks (Continued)

To open . . .	Select . . .
Count Groups	Scripts > Basic Scripts > Numbers > Count Groups
Clustering	Scripts > Basic Scripts > QC & Analysis > Clustering
Correlations	Scripts > Basic Scripts > QC & Analysis > Correlations
Filters	Scripts > Basic Scripts > QC & Analysis > Filter Gene Tools
Regulatory Sequences	Scripts > Basic Scripts > QC & Analysis > Regulatory Sequence Search
Statistical Analysis	Scripts > Basic Scripts > QC & Analysis > Statistical Analysis (ANOVA)
Standardize Inputs	Scripts > Basic Scripts > Standardize Inputs
External Programs	Scripts > External Programs

Using Building Blocks from External Programs

External program building blocks are points of contact for other programs. External program building blocks are not editable, but can be used in any script. Typically, these external program building blocks appear in the Script Editor's Navigator panel. If you do not see any, use this procedure to access the external program.

To access a building block from an external program:

1. Select **Tools** > **Script Editor**.
2. Click the **File Access** link to download the samples from the following web site:

   `http://www.signetics.com/cgi/SiG.cgi/Products/GeneSpring/extProgs.smf`.

3. Place the jar file in the folder `..genespring\data\programs\`.

You may have more external program building blocks, as GeneSpring and Script Editor will create an external program building block for every external program in your GeneSpring. You may not have any external program building blocks in the Script Editor if you are working on an older version on GeneSpring, or if you do not have any external programs.

Managing Scripts

A typical script is built by arranging building blocks in the browser. Building blocks can be other scripts, script building blocks or external program building blocks. This section explains how to build scripts. It covers the following topics:

- Creating Scripts
- Defining Knobs
- Converting Knobs to Input
- Saving a Building Blocks as a Script
- Arranging Inputs and Outputs in the Browser
- Dynamically Naming Script Outputs
- Saving Scripts
- Saving Scripts to Subfolders
- Moving Scripts
- Warning Messages
- Getting Help with Scripts

Creating Scripts

To create a script:

1. Select **Tools** > **Script Editor**.
2. In the **Navigator**, locate the script you want.
3. Drag the building block you want to the **Browser**.
4. Click and drag the edges of the building block to make it larger or smaller.

   Details about the building block appear in the **Block** section on the right side of the main Script Editor window as long as it is selected. You can click anywhere else in the browser to de-select the building block.
5. (Optional) Define or expose the knobs you want.

   Note: If you don't specify or expose any knobs, the defaults will be used.

   Go to "Defining Knobs" on page 16-41 for instructions

   These can be menus or text fields, and appear in the lower right section of the Script Editor window.
6. Create a line from the input (top) to the first building block.
7. Select a new building block from the folders in the Navigator.
8. Connect the first building block to the second.
9. Repeat steps 2 through 6 as necessary.

10. When you are close to done, check the error messages at the bottom to make sure everything is connected properly.
11. Name and save your script.

    Go to “Saving Scripts” on page 16-45 for instructions.
12. If you would like to create a new script immediately, select **File > New**.

    You can make several small scripts and join them together.

Defining Knobs

To define a knob:

1. In the **Navigator**, select the building block you want.
2. Right-click in the **Knobs** area and select **Add Knob** from the menu.
3. Select the appropriate variable type. The available types are:
   - Integer
   - Positive Integer
   - Number
   - Positive Number
   - Yes or No
   - Measurement Type
   - Percentage
   - Correlation
   - Name
   - Comparison
   - Chromosome Number
   - Gene Annotations
4. (Optional) Enter a default value for the knob in the Default Value field in the **Notes** field.

Converting Knobs to Input

You can convert certain types of knobs and use them as input for a script (Figure 16-13). This knobs include the following:

- Boolean
- Integer number
- Positive integer
- Positive number
- Percentage

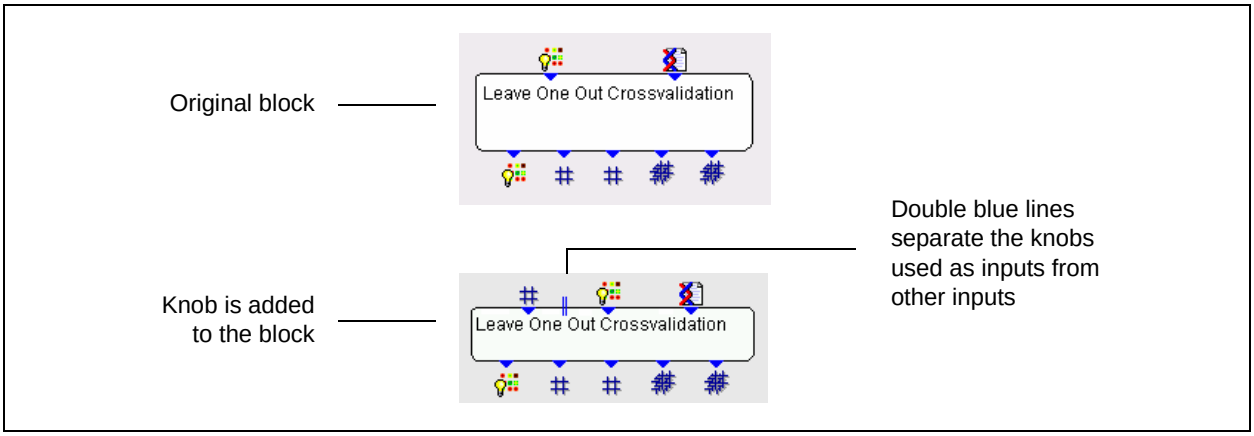
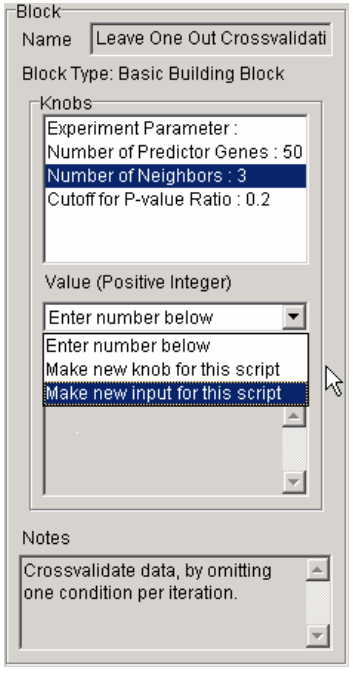

Original block
Leave One Out Crossvalidation
Double blue lines separate the knobs used as inputs from other inputs
Knob is added to the block
Leave One Out Crossvalidation

Figure 16-13 Example Knob as Input

To convert a knob to input:

1. In the **Navigator**, select the building block you want.
2. In the **Knobs** box, select the knob you want.
3. In the **Value** menu, select the **Make new input for this script** option.



The knob automatically appears at the top of the block. It is now available to function as input for a script.

Converting Input Back to Knobs

This procedure explains how to convert input back into a knob. Only those inputs that were upgraded from knob can be downgraded to a knob.

To convert the input back to a knob:

1. In the **Navigator**, select the building block that contains the input promoted from a knob.
2. In the right panel, click on the knob that got promoted.
3. In the **Value** menu, select any option other than **Make new knob for this script**.

Saving a Building Blocks as a Script

You can save a block built from multiple other blocks as a script.This is useful for updating obsolete script building blocks, or if you want to use a block from a script someone else created.

To save a building block as a script:

- Right-click the building block and select the **Save Block as Script** option.

Arranging Inputs and Outputs in the Browser

When you right-click over a socket icon in the browser, a menu appears. Which options appear depends on the status of the selected socket.

- **Move to Start**—This command moves the socket all the way to the left of the inputs or outputs field. It lets you “untangle” the web of connecting lines.
- **Move to End**—This command moves the socket all the way to the right of the inputs or outputs field. It lets you “untangle” the web of connecting lines.
- **Remove Node**—You will not get a warning message before the socket or node is removed from your potential script.

A number appears in parentheses after each socket in the input field. The numbers represent the order in which the input will be presented when the script is run.

Dynamically Naming Script Outputs

You can define output names dynamically based on the value of a script input, parameter, or note, as follows:

```
$name$
$param$
$notes$
```

Note: These words are case sensitive.

For example, the sample script *Best k-means* has two inputs, a gene list and an experiment. If you want the script output to include the name of the gene list and the experiment, do the following:

1. Select the appropriate script output socket.
2. In the **Make Name** field in the Notes section of the Script Editor, enter the following:

```
Best k-means $name$1 in $name2
```

3. Save the script.

If you run the *Best k-means* script using the gene list "ACGCGT in all ORFs" and the experiment "Extraterrestrial Yeast Study", the output classification is automatically named "Best k-means ACGCGT in all ORFs in Extraterrestrial Yeast Study".

The procedure is similar when basing output names on knob values, except that the format for knob values is "`$param$x`", where *x* is the number of the knob. So, for example, to cause the output of the Filter on Noise sample script to include the name of the input experiment and the value of the Filter Cutoff knob, enter the following in the **Make Name** field:

```
Filter on Noise on $name$1 with Filter cutoff $param$1
```

If your input experiment is "Extraterrestrial Yeast Study" and the filter cutoff is "0.1", the output gene list is named "Filter on Noise on Extraterrestrial Yeast Study with Filter cutoff 0.1".

Any character that is legal in the name of a Navigator object is legal in the **Make Name** field. The only illegal characters are the forward slash (/) and back slash (\).

Creating Useful Notes for Output Objects

The same mechanism that is used to give script outputs useful names, can be used to fill the notes section for each of the objects with useful information. The Notes section for a gene list could be filled with a textual description of how the gene list was created, what the script settings were, and what the names of the input objects were, etc. For example:

```
This gene list was created by the Volcano script and contains
the downregulated genes.

Script settings:
============
Starting Genelist: "$name$2"
Condition: "$name$1"
P value cutoff: "$param$3"
Fold change cutoffs" $param$1" and "$param$2"
```

These settings will fill the Notes section of the resulting gene list with notes that are sufficiently detailed as to allow the user to determine exactly how the gene list was created.

The special variable names like `$name$1` and `$param$2`, etc. will be replaced with the names and settings of the actual inputs and knobs, as described in "Dynamically Naming Script Outputs" on page 16-43.

To edit the notes for a output object:

1. In the Script Editor, select the output object you want.
2. In the Top Notes section, enter information about the output object.

   These notes, which can be used to inform developers about the function of the output object, are only visible in the Script Editor.
3. In the Bottom Notes section, enter the information you want to associate with the notes for the resulting output object.

In Figure 16-14, the Bottom Notes section shows the notes that will be copied to the output gene list. The Top Notes section is only used for documentation purposes.

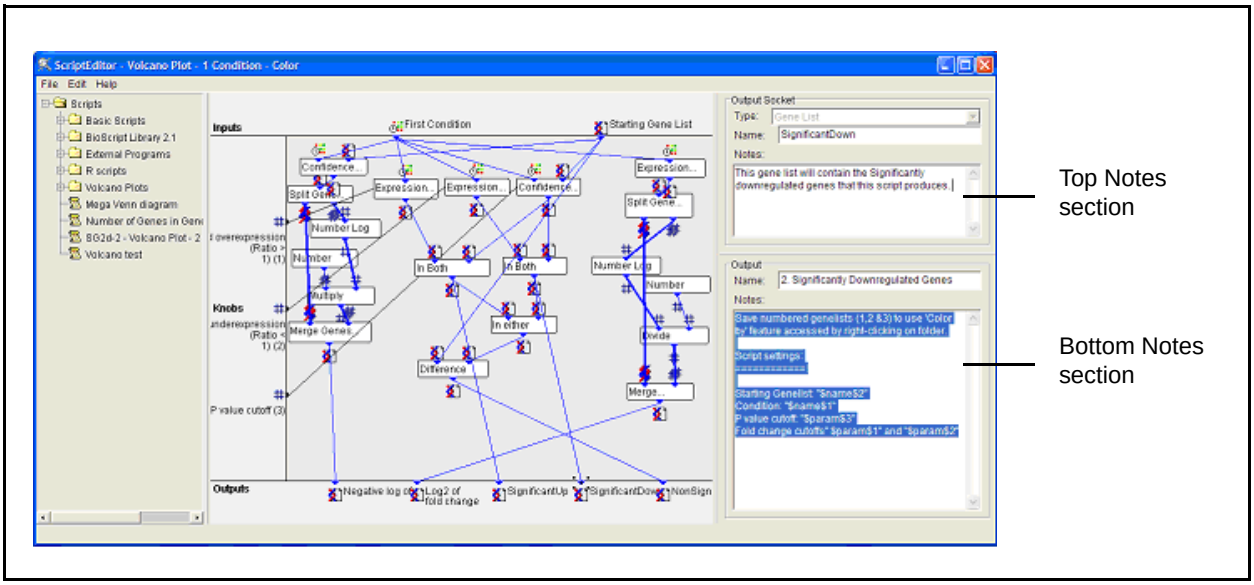


Figure 16-14 Script Editor: Notes

Saving Scripts

You can save either finished or unfinished scripts. Unfinished scripts are displayed using a red icon in the Navigator.

To save a script:

- Select **File > Save**.

The first time you save a script, a **Save As** window opens in which you can specify a name for your script and a folder in which to save it. If an error message appears saying your result cannot be saved, rename the script and try again.

Saving Scripts to Subfolders

By default, the Script Editor saves scripts in the main Scripts folder. If you want, you can also save the script in a subfolder.

To create a subfolder within a folder:

- Do one of the following:
  - Right-click the parent folder in the **Navigator**, and select **Add Folder**.
  - If you enter a nonexistent folder name in the **Save As** window, the Script Editor creates the directory for you and saves your script in it.

Moving Scripts

To move a script:

- Do one of the following:
  - Drag the script from the **Navigator** and drop it to the location you want.
  - Right-click the script and select the **Move** command.

Warning Messages

If GeneSpring detects any problems or missing information in your script, warning or error messages may appear across the bottom of the window.

- Warning messages are always preceded by the word *Warning* and are displayed in orange text.
- Error messages appear in dark red text.

Notes:

- You can save a script when a warning message is active, but it may not perform as expected.
- You cannot run a script with an active error message.
- Scripts with active error messages cannot be used as building blocks

Getting Help with Scripts

Live technical support for all products is available from 7 A.M. - 5 P.M. Pacific Standard Time. Please contact us at techsupport@silicongenetics.com or call us at +1-(866) SIG SOFT(866) 744 7638 (in the US) or +81 (70) 5072-2085 (Japan) or +44 (0) 1259 751 833 (Europe). Support via email is also available at: support@silicongenetics.com.

Working with External Programs

This section explains how to use external program in GeneSpring. It covers the following topics:

- External Program Interface
- External Programs Folder
- External Program Script Building Blocks
- New External Program Window
- Creating External Programs
- Running External Programs
- External Program Example
- Inspecting External Programs

External Program Interface

The GeneSpring External Program interface allows you to run external analysis programs from within GeneSpring. These programs can be useful when your research calls for a type of analysis that GeneSpring does not perform. The external program interface is also useful for parsing and pre-formatting data for use in another application.

When you launch an external program from within GeneSpring, the data that is displayed in the Genome Browser will be sent to the external program as standard input. When the external program runs, GeneSpring recognizes the standard output generated by the external program and displays it in the Genome Browser.

External Programs Folder

External programs are listed in the External Programs folder in the Navigator. Each external program is also automatically wrapped into an external program script building block so it can be used to build scripts. These building blocks are shown in the GeneSpring Navigator (Scripts > External Programs folder) and in the Script Editor's Navigator (External Programs folder). In these folders, you will find an external program building block for each external program defined in GeneSpring.

External Program Script Building Blocks

There are two significant types of script building blocks for external programs:

- **Load data object to file**—The loading external program building block makes a specified data object from outside GeneSpring available for use in a script.
- **Save data object to file**—The saving external program building block saves the output from a script to your local hard drive. This script does not save the data object to GeneSpring or Signet; there are other building blocks that perform these functions.

GeneSpring checks for new, changed, or deleted external programs each time it is started.

If you have shared a script that contains an external program script building block with a colleague, ensure that your colleague has access to all external programs needed to run a script.

If an external program is required to run a script, and that external program is not available on a local version of GeneSpring, an error message results in the Run Script window (Figure 16-15).

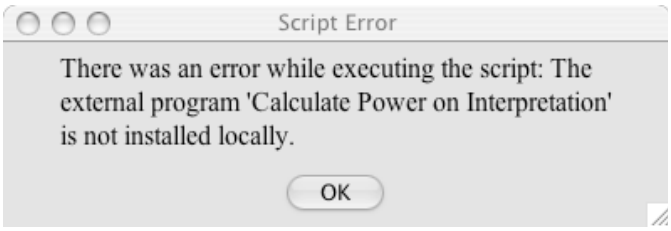

Script Error
There was an error while executing the script: The external program 'Calculate Power on Interpretation' is not installed locally.
OK

Figure 16-15 Script Missing External Program Error

It is still possible to store and share scripts with missing external programs using Signet, but you must ensure that the corresponding external programs are installed on all GeneSpring installations on which they will run.

To outsource computations of scripts that use external programs to a remote server, you must ensure that the external programs are installed on all remote servers on which they will run. Additionally, you must ensure that the external programs you choose to run don't require a graphical user interface (because the GUI cannot be shown on a remote server).

New External Program Window

The New External Program window (Figure 16-17) lets you install an external program that you want to run in GeneSpring. From this window, you can specify the inputs, outputs, program type, and other information about an external program. Any program capable of receiving standard input, or producing data on the standard output, can be run directly from GeneSpring.

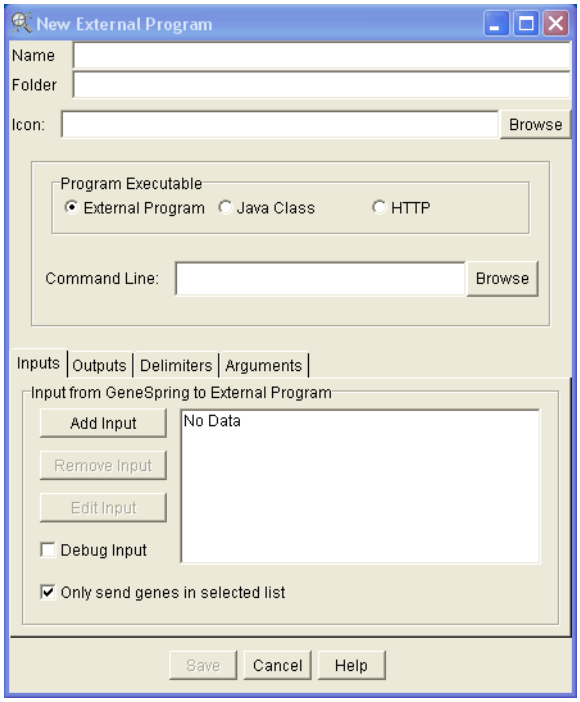


Figure 16-16 New External Program Window

This window contains the following elements:

- **Name**—The name of the program as it will appear in the GeneSpring Navigator. Choose a descriptive name that will be easy to remember.
- **Folder**—The folder in which to save the external program definition.
- **Icon—**The gif file you want to use to represent this external program. You can use the Browse button to locate the image file.
- **Program Executable**—The type of program (External program, Java class, or HTTP). Depending on the option selected, you will have to specify the following:
  - **External Program**—The complete path name required to run the program.
  - **Java Class**—The full Java class name and the location of the corresponding jar file.
  - **HTTP**—The URL link.
- **Command Line**—The complete path name required to run the program.
- **Inputs**—Data input for the external program.
- **Outputs**—Output of the external program.
- **Delimiters**—Delimiters to separate multiple outputs.
- **Arguments**—Optional command line arguments for the external program.

Creating External Programs

This section explains how to create external programs. It covers the following topics:

- Creating New External Programs
- Adding Inputs to External Programs
- Adding Outputs to External Programs
- Adding Custom Delimiters
- Adding Arguments
- Removing Arguments

Creating New External Programs

To create a new external program:

1. Select **File > New Program > New External Program**.

   The New External Program window opens.

2. In the **Name** field, enter the name of the program as it will appear in the GeneSpring Navigator.

3. In the **Folder** field enter the name of the folder in which to save the external program definition.

   - To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the window. The selected folder appears in the **Folder** field.
   - To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.

4. (Optional) In the **Icon** field, **click** Browse, and locate the gif file you want to use to represent this external program.

   If you do not provide an icon, a generic image will be used.

5. In the **Program Executable** box, select the program type (external program, Java class, or HTTP link).

6. In the **Command Line** field, enter the complete path name required to run the program. For example:

   ```
   ..\Perl\ScriptCompletionHaiku.pl or
   http://ecoli.sample.com/analysis.cgi.
   ```

7. Do one or more of the following:

   - To add inputs, go to “Adding Inputs to External Programs” on page 16-51.
   - To add outputs, go to “Adding Outputs to External Programs” on page 16-52.
   - To add delimiters, go to “Adding Custom Delimiters” on page 16-54.
   - To add command-line arguments, go to “Adding Arguments” on page 16-54.

Adding Inputs to External Programs

The input is what GeneSpring sends to the external program. On this tab, specify any necessary input data for the external program. You can add as many inputs as you like.

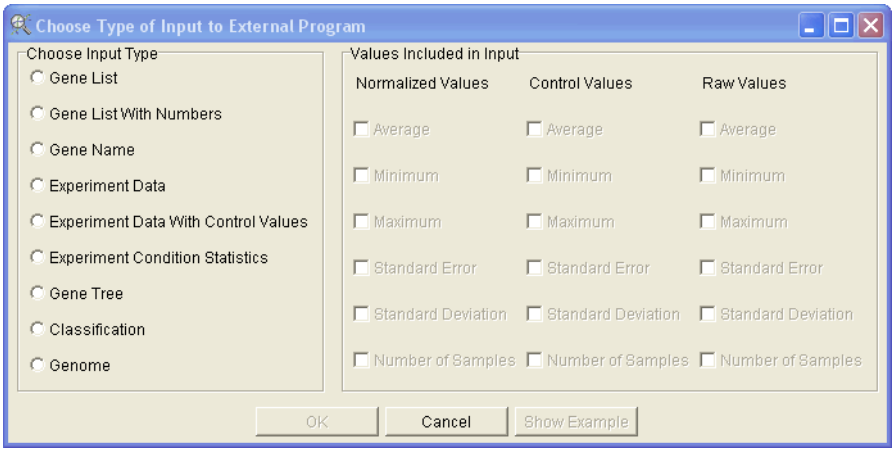
To add inputs:

1. Select **File > New Program >New External Program**.

   The New External Program window opens.

2. Click **Add Input**.

   The Choose Type of Input window opens.



3. Select the type of input to send to the external program. The available choices are:

   - **Gene List**—A tab-delimited list of systematic names, one per line
   - **Gene List With Numbers**—A tab-delimited list of systematic names and associated numbers, one pair per line
   - **Gene Name**—A single systematic name
   - **Experiment Data**—Normalized experiment data, one line per gene, one column per experiment, with header lines for the experiment name and each parameter. Only genes in the currently selected gene list are sent.
   - **Experiment Data with Confidence—**Normalized experiment data, one line per gene, two lines per experiment (one for normalized data and one for confidence values), with header lines for the experiment name and each parameter. Only genes in the currently selected gene list are sent.
   - **Experiment Condition Statistics**—One gene per line, one column per statistical quantity per condition, with labels for the statistics across the top in the form `{N,R,C}_{AVERAGE, MIN, MAX, STDERR, STDDEV, N}`.
   - **Classification**—A tab-delimited list of systematic names and the name of the associated classification group, one pair per line. Only genes in the currently selected gene list and classification are sent.
   - **Gene Tree—**A hierarchical tree in XML format.

- **Genome**—An XML representation of the genome, which will be automatically saved to disk during Import to GeneSpring.

Note: The Values Included in Input panel is active only if you selected the Experiment Condition Statistics option.

4. Click **Show Example** for an example of the selected output. Examples can be viewed as plain text, hex code, or as a spreadsheet. Characters that cannot be displayed, such as tabs, appear as boxes.
5. Click **OK** to return to the New External Program window.
6. (Optional) Select the **Debug Input** option.

   This option writes to the console window when input is sent from GeneSpring to the external program.
7. To edit an existing input type, select it in the list box and click **Edit Input**.
8. To remove an input type, select it in the list box and click **Remove Input**.

Adding Outputs to External Programs

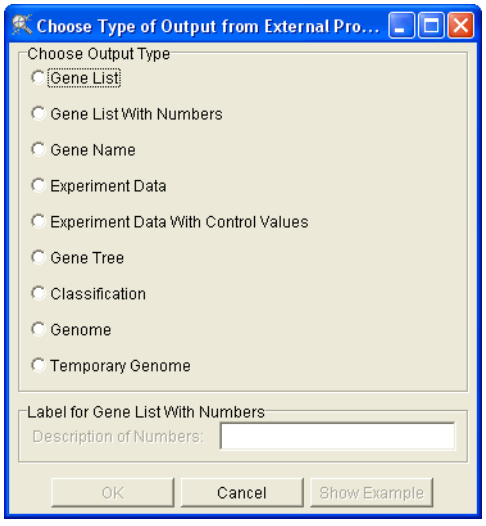
The output is what GeneSpring receives from the external program. On this tab, specify the desired output of the external program. You may add as many outputs as you like. If the external program does not send any data back to GeneSpring, you do not need to enter anything in this tab.

To add an output:

1. Select **File > New Program >New External Program**.

   The New External Program window opens.
2. Click the **Outputs** tab.
3. Click **Add Output**.

   The Choose Type of Output window opens.



4. Specify the type of output GeneSpring will receive from the external program. The available options are:
   - **Gene List**—A tab-delimited list of systematic names, one per line
   - **Gene List With Numbers**—A tab-delimited list of systematic names and associated numbers, one pair per line
   - **Gene Name**—A single systematic name
   - **Experiment Data**—Normalized experiment data, one line per gene, one column per experiment, with header lines for the experiment name and each parameter
   - **Experiment Data with Control Values**—Normalized experiment data, one line per gene, two lines per experiment (one for normalized data and one for control values), with header lines for the experiment name and each parameter
   - **Gene Tree**—A hierarchical tree in XML format
   - **Classification**—A tab-delimited list of systematic names and the name of the associated classification group, one pair per line
   - **Genome**—An XML representation of the genome, which will be automatically saved to disk during Import to GeneSpring
   - **Temporary Genome**—Identical to the Genome format, except that it is not saved to disk
5. Click **Show Example** for an example of the selected output.

   Examples can be viewed as plain text, hex code, or as a spreadsheet. Characters that cannot be displayed, such as tabs, appear as boxes.
6. If the output is a gene list with numbers, enter a label to describe what the numbers represent in the **Description of Numbers** field.

   This label appears in the Gene List Inspector as the title of the column of numbers.
7. Click **OK** to return to the New External Program window.
8. (Optional) Select the **Debug Output** option.

   This option writes to the console window when output is sent from the external program to GeneSpring.
9. To edit an existing output type, select it in the list box and click **Edit Output**.
10. To remove an output type, select it in the list box and click **Remove Output**.

Adding Custom Delimiters

Both GeneSpring and the external program need to know when a new data type is being sent. Certain characters are used to indicate this new data type. This is usually the ASCII 255 character, however, your program may require a different delimiter.

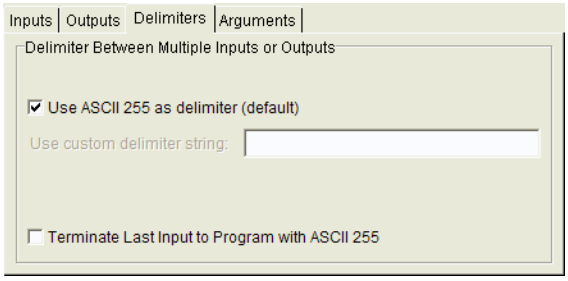
By default, GeneSpring uses ASCII 255 as the data type delimiter.

To enter a custom delimiter:

1. Select **File > New Program >New External Program**.

   The New External Program window opens.

2. Click the **Delimiters** tab.



3. Clear the **Use ASCII 255 as delimiter** box.
4. In the **Use custom delimiter string** text box, enter the delimiter you want.

   You can enter a character or a string.

5. Some external programs may look for ASCII 255 to indicate that the data has finished being sent. If this is the case, select the **Terminate Last Input to Program with ASCII 255** option.

Adding Arguments

Command line arguments are a way of providing extra information to the external program. For example, if the external program can perform one of three clustering methods, a command line argument might tell the external program which clustering method to use.

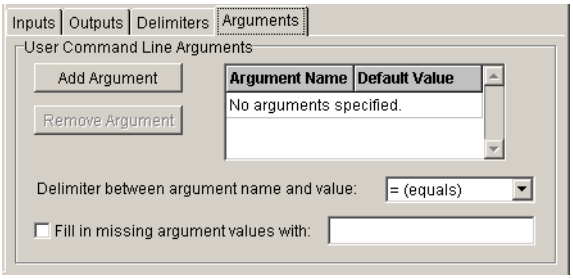
This procedure is optional since only some external programs require command line arguments.

To add an argument:

1. Select **File > New Program >New External Program**.

   The New External Program window opens.

2. Click the **Arguments** tab.



This tab contains a table of name-value pairs displaying the name of the argument and its default value.

3. Click **Add Argument**.

   In the table to the right, a new line appears.

4. Replace the text `name1` and `value1` with the appropriate argument name and value; for example: `-v` and `all`.

   Some arguments may not have values. In this case, enter only the argument name, and delete the sample text `value1` and leave the Default Value field blank.

5. Select the separator between the argument name and the argument value. This is either `=` or `:`.

6. Some versions of Windows will not correctly match up argument name/value pairs, and will read values as new arguments. To avoid this problem, check the **Fill in missing argument values** box and enter a filler term in the provided text box.

   This filler term should be something the external program will ignore, or a character that doesn't occur normally, such as ASCII 255.

7. Click **Save**.

Removing Arguments

1. Select **File > New Program >New External Program**.

   The New External Program window opens.

2. Click the **Arguments** tab.

3. Select the argument you want to remove.

4. Click **Remove Argument**.

Running External Programs

The GeneSpring External Program interface lets you run external analysis programs from within GeneSpring. These programs can be useful when your research calls for a type of analysis that GeneSpring does not perform. The external program interface is also useful for parsing and pre-formatting data for use in another application.

When you launch an external program from within GeneSpring, the data that is displayed in the Genome Browser is sent to the external program as standard input. When the external program runs, GeneSpring recognizes the standard output generated by the external program and displays it in the Genome Browser.

To run an external program:

1. Do one of the following:
   - In the **Navigator**, select the **External Programs** folder, and click the external program you want.
   - In the **Navigator**, select the **External Programs** folder, right-click the external program you want, and then select the **Run** command.

External Program Example

This example demonstrates how to use GeneSpring's external program interface. The External Program Interface exports GeneSpring experimental data, runs a SAS™ program to analyze it, and brings the results back into GeneSpring for display. This example was developed with Windows 2000 using SAS™ version 8. It should work with earlier versions of Windows, but earlier versions of SAS™ require some modifications.

This particular example sets up an interface to the SAS™ procedure FASTCLUS to do gene clustering. You will need to create two text files with an ASCII text editor such as Notepad. These files are *runsas.bat* and *fastclus.sas*.

1. Create a batch file called *runsas.bat*.

   This batch file takes the standard input from GeneSpring, stores it in a file, executes SAS™, and passes the results back to GeneSpring via standard output. The program *cat.exe* simply copies standard input into standard output. If you do not have something equivalent on your system, `cat.exe` can be downloaded from the Silicon Genetics web site.

2. Insert the following text in the batch file:

```
@echo off
set infile=%2
set outfile=%3
cat.exe > %2
set SASROOT=C:\PROGRA~1\SASINS~1\SAS\V8
%SASROOT%\SAS %1.sas -nologo -config %SASROOT%\SASV8.CFG
cat.exe < %3
del %1.lst %1.log %2 %3
```

3. Create a text file called *fastclus.sas*.

This batch file runs PROC FASTCLUS, specifying 5 clusters. In PROC IMPORT, the `datarow=3` command skips the first two lines of the exported data, which contain the dataset name and one parameter. If you have more than one parameter, adjust the datarow value accordingly.

PROC EXPORT puts a header line on the return data set listing the variable names, and GeneSpring displays an error message and skips this line (unless you have a gene named VAR1, in which case you should rename VAR1 to something else in your application).

4. Insert the following text in the batch file:

```
filename infile "%sysget(infile)";
filename outfile "%sysget(outfile)";

proc import datafile=infile DBMS=TAB out=experiment replace;
datarow=3;
getnames=no;
run;

proc fastclus data=experiment maxclusters=5 maxiter=50
out=clusters(keep=var1 cluster);
id var1;
run;

proc export data=clusters outfile=outfile DBMS=TAB replace;

run;
```

5. Save the batch file.
6. Select File > New External Program.

   The New External Program window opens.
7. Enter "SAS FASTCLUS" in the Name field.
8. Leave the **Folder** field blank.

   The external program is saved in the External Programs folder by default.
9. Select the **External Program** option.
10. Enter the following text in the **Command Line** field:

```
runsas.bat fastclus expt.txt clus.txt
```

11. On the **Inputs** tab, click **Add Input** and select the **Experiment Data** option.
12. On the Outputs tab, click **Add Output** and select the **Experiment Data with Control Values** option.
13. Click the **Save** button.

Your external program should now appear in the GeneSpring Navigator in the External Programs folder. The corresponding external program script building block appears in the Scripts > External Programs folder in the Navigator (External Programs in ScriptEditor).

Inspecting External Programs

You use the External Program inspector to view and modify details relating to an external program. This section covers the following topics:

- External Program Inspector Window
- Opening the External Program Inspector
- Editing External Programs
- Data Formats for the External Program Interface

External Program Inspector Window

The External Program Inspector window (Figure 16-17) lets you view details of an existing external program.

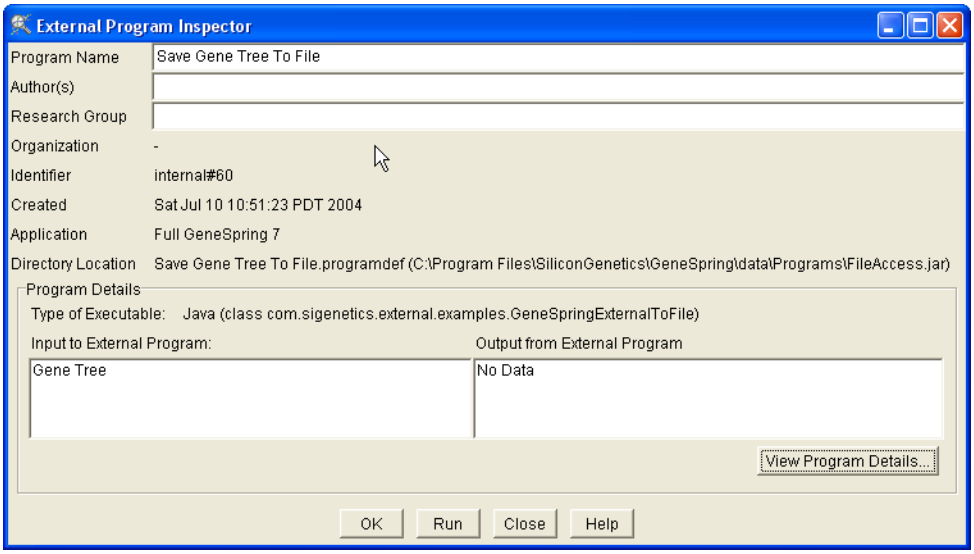


Figure 16-17 External Program Inspector Window

Summary Information Section

The top portion of the window displays the following information:

- **Name**—Show the name of the external program.
- **Author(s)**—Lets you enter the name of the authors associated with the external program.
- **Research Group**—Lets you enter the name of the research group associated with the external program.
- **Organization**—Shows the name of the organization associated with the external program. This information cannot be edited; it is derived from the license key.
- **Created**—Shows the date and time at which the external program was created.

- **Application**—Shows the GeneSpring name and version number used to import the external program.
- **Location**—Shows the name of the directory in which the external program is saved.
- **Notes**—Lets you enter information about the external program.

Program Details Section

The Program Details panel contains the following information:

- **Type of Executable**—Shows the executable type, for example: Java class, external program, or URL.
- **Command**—Shows the command line (visible only if the external program is a command line executable).
- **Input to External Program**—Shows the inputs defined for the external program.
- **Output from External Program**—Shows the outputs defined for the external program.
- **View Program Details**—Opens the View External Program window.

Opening the External Program Inspector

To open the External Program Inspector:

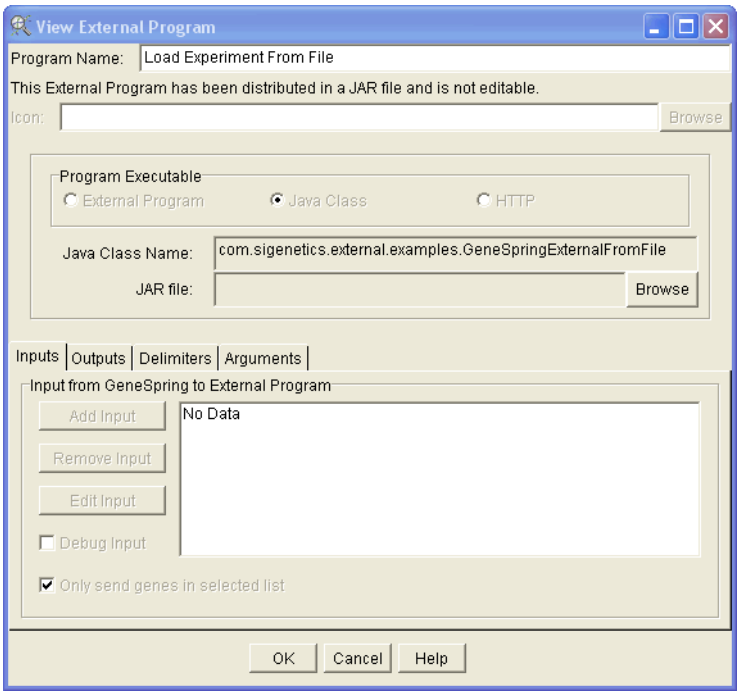
1. In the **Navigator**, select the **External Programs** folder.
2. Right-click the external program you want and select the **Inspect** command.

Editing External Programs

To open the External Program Inspector:

1. In the **Navigator**, select the **External Programs** folder.
2. Right-click the external program you want and select the **Inspect** command.

   The External Program Inspector window opens.
3. Edit the **Program Name**, **Author**, and **Research Group** fields.
4. Do one of the following:
   - If a program cannot be edited, such as the external programs included with GeneSpring, the button in the lower right portion of the window is labeled **View Program Details**. Click this button to view additional details about the program. You cannot make any changes from the View Program Details window.



- If the program can be edited, the button displays **Edit Program Details**. To edit the program, click **Edit Program Details** button.

This window is identical to the Create New External Program window. See "New External Program Window" on page 16-48 for more information.

5. Click **OK**.

Data Formats for the External Program Interface

You can choose the data to send to an external program, as well as the format in which the data is sent and received. All formats are optionally terminated in a special termination character, ASCII 255. The formats are generally text files.

You can also send or receive multiple data objects. For example, your program might want to receive both the currently selected gene list and the currently selected experiment; or it might send a new genome to GeneSpring, followed by an experiment for that genome.

Table 16-23 describes the data formats for the external program interface.

Table 16-23 External Program Interface—Data Formats

Format	Input	Output	Num	Description	Example
No data	Yes	Yes	0	No data. Typically used for one way communication.	
Gene List	Yes	Yes	1	List of gene names, one per line	L20294 M89777 X95403 M63630
Gene List (with numbers)	Yes	Yes	2	List of gene names, one per line, with each gene followed by a tab and an associated number	L20294 1 M89777 3 X95403 3.566 M63630 0
Gene Name	Yes	Yes	3	The name of one gene	L20294
Experiment Data	Yes	Yes	4	Experimental data. One line per gene, one column per sample, with a header line for each parameter.	Tup1 deletion experiment time (minutes) 1 2 YPR1 0.88 1.09 YGR1 1.81 1.63 YNL1 0.52 1.18
Experiment data with confidence	Yes	Yes	5	As above, except two columns per sample. The first column has normalized data, the second has confidence values.	Tup1 deletion experiment time (minutes) 1 control 2 control YPR1 8 43 9 70 YGR1 1 7.3 3 7 YNL1 5 4.9 8 49
Classification	Yes	Yes	6	One gene per line, followed by tab, followed by name of classification	146 set3 158 unclassified 159 set5 170 set3 171 set3 181 set1

Table 16-23 External Program Interface—Data Formats (Continued)

Format	Input	Output	Num	Description	Example
Tree	Yes	Yes	7	Hierarchical	<TREE DISTANCE=0.6 TITLE=a> YMR199W YPL256C <TREE DISTANCE=0.1 TITLE=b> YAL001C YAL002W </TREE> <TREE DISTANCE=0.2 TITLE=c> YAL019W YAL017W </TREE> </TREE>
Genome	Yes	Yes	8	XML description of genome. Hyperlinks and sequence are optional.	<GENOME CIRCULAR="false"> <NAME>Rat</NAME> <HYPERLINKS> %GenBank;http://www.ncbi.nlm.nih.gov/... $PubMed;http://www.ncbi.nlm.nih.gov:80/... </HYPERLINKS> <MAPPED_FILE> GAD65 GAD2 4.1.1.15 glutamic acid decarboxylase ... pre-GAD67 GAD67 4.1.1.15 glutamic acid decarboxylase ... ... </MAPPED_FILE> <SEQUENCE> >CHR1 Chromosome I data: CCACACCACACCCACACACCCACAC... </SEQUENCE> </GENOME>
Temporary Genome	Yes	Yes	9	Same as above, but genome is not saved to disk. No data for it can be saved, and all records of the genome disappear when the window is closed.	Same as above

Working with Plugins

A plugin is an external program that lets you perform tasks within GeneSpring. This section explains how to install and run plugins. It covers the following topics:

- Plugin Types
- Working with Data Preprocessor Plugins
- Working with Interactive Plugins
- Working with Script Plugins

Plugin Types

GeneSpring enables you to create the following types of plugins:

- **Data Preprocessor Plugin**—Lets you apply arbitrary transformations on data files (for example, the default normalization algorithm), before importing the data into GeneSpring.
- **Interactive Plugin**—Lets you write Java programs that add features to GeneSpring.
- **Script Plugin**—Lets you invoke an arbitrary computation as a script building block. In order to create a script plugin, you will have to define all the parameters for running a script interactively. The parameters include the script inputs, knobs, outputs, and the specification of the program that performs the computation.

Working with Data Preprocessor Plugins

Data preprocessors let you create your own data pre-processing program, such as RMA, to apply a default normalization algorithm to data during the import process. These preprocessors enable GeneSpring to recognize currently unrecognizable file formats, such as Affymetrix *cel* files. The Data Preprocessors show up in GeneSpring only when you attempt to load files that are readable in a format that a preprocessor recognizes.

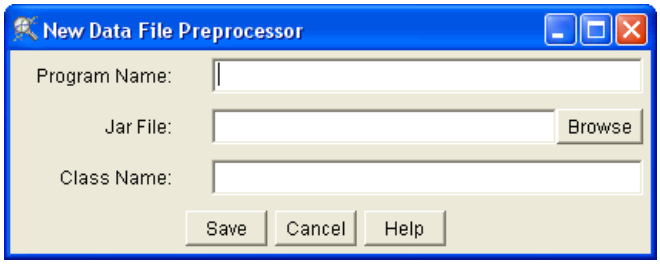
This section explains how to install a data preprocessor in GeneSpring. It does not describe how to build a data preprocessor. For more information on building a plugin, go to the *GeneSpring API JAVADOC*. To access this document, go to the API folder in the GeneSpring doc directory and open the index.html file.

Installing Data Preprocessor Plugins

To install a data preprocessor plugin:

1. Select **File > New Program > New Data Preprocessor Plugin**.

   The New Data Preprocessor Plugin window opens.



2. In the **Program Name** field, enter the name of the external program as it will appear in the **Navigator**.
3. In the **Jar File** field, click **Browse**, and select the full path name to the jar file that contains the program you want to run.
4. In the **Class Name** field, enter the fully-qualified class name (which includes the dots).
5. Click **Save**.

Working with Interactive Plugins

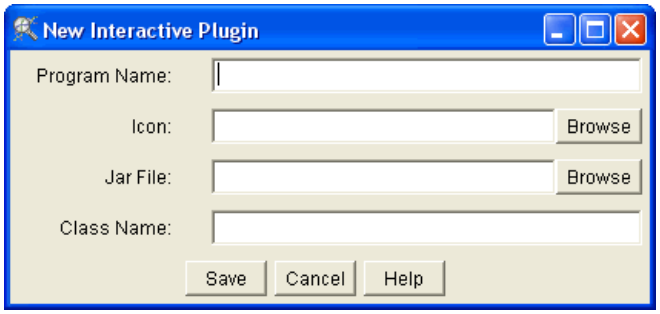
This section explains how to install, run, and edit an interactive plugin in GeneSpring. It does not describe how to build an interactive plugin. For more information on building a plugin, go to the *GeneSpring API JAVADOC*. To access this document, go to the API folder in the GeneSpring doc directory and open the index.html file.

Installing Interactive Plugins

To create an interactive plugin:

1. Select **File > New Program > New Interactive Plugin**.

   The New Interactive Plugin window opens.



2. In the **Program Name** field, enter the name of the external program as it will appear in the **Navigator**.
3. (Optional) In the **Icon** field, **click** Browse, and locate the gif file you want to use to represent this external program.

   If you do not provide an icon, a generic image will be used.

4. In the **Jar File** field, do one of the following:
   - Enter the full pathname to the jar file that contains the program you want to run.
   - Click **Browse** and select the full path name to the jar file you want.
5. In the **Class Name** field, enter the fully-qualified class name (which includes the dots).
6. Click **Save**.

   The External Program Inspector window opens.
7. Go to "Running Interactive Plugins" on page 16-65.

Running Interactive Plugins

After installing and saving an interactive plugin, GeneSpring displays the External Program Inspector window.

To run an interactive plugin:

- Do one of the following:
  - In the **Navigator**, open the **External Program** folder, and click the interactive plugin you want to run.
  - In the **Navigator**, select the **External Programs** folder, right-click the interactive plugin you want, and then select the **Inspect** command. In the External Program Inspector window, click the **Run** button.

Editing Interactive Plugins

To edit an interactive plugin:

1. In the **Navigator**, select the **External Programs** folder, right-click the interactive plugin you want, and then select the **Inspect** command.

   The External Program Inspector window opens.
2. Edit the **Program Name**, **Author**, and **Research Group** fields.
3. Click **Edit Program Details**.

   The Edit Interactive Plugin window opens.
4. In the **Program Name** field, enter the name of the external program as it will appear in the **Navigator**.
5. In the **Icon** field, do one of the following:
   - Enter the full pathname to the gif file you want to use to represent this interactive plugin.
   - Click **Browse** and locate the gif file you want.
6. In the **Jar File** field, do one of the following:
   - Enter the full pathname to the jar file that contains the program you want to run.
   - Click **Browse** and select the full path name to the jar file you want.

7. In the **Class Name** field, enter the fully-qualified class name (which includes the dots).
8. Click **Save**.

Working with Script Plugins

A Script plugin lets you invoke an arbitrary computation as a script building block. In order to create a script plugin, you will have to define all the parameters for running a script interactively. The parameters include the script inputs, knobs, outputs, and the specification of the program that performs the computation.

New Script Plugin Window

The New Script Plugin window (Figure 16-18) lets you specify parameters for running a script interactively. The parameters include the script inputs, knobs, and outputs.

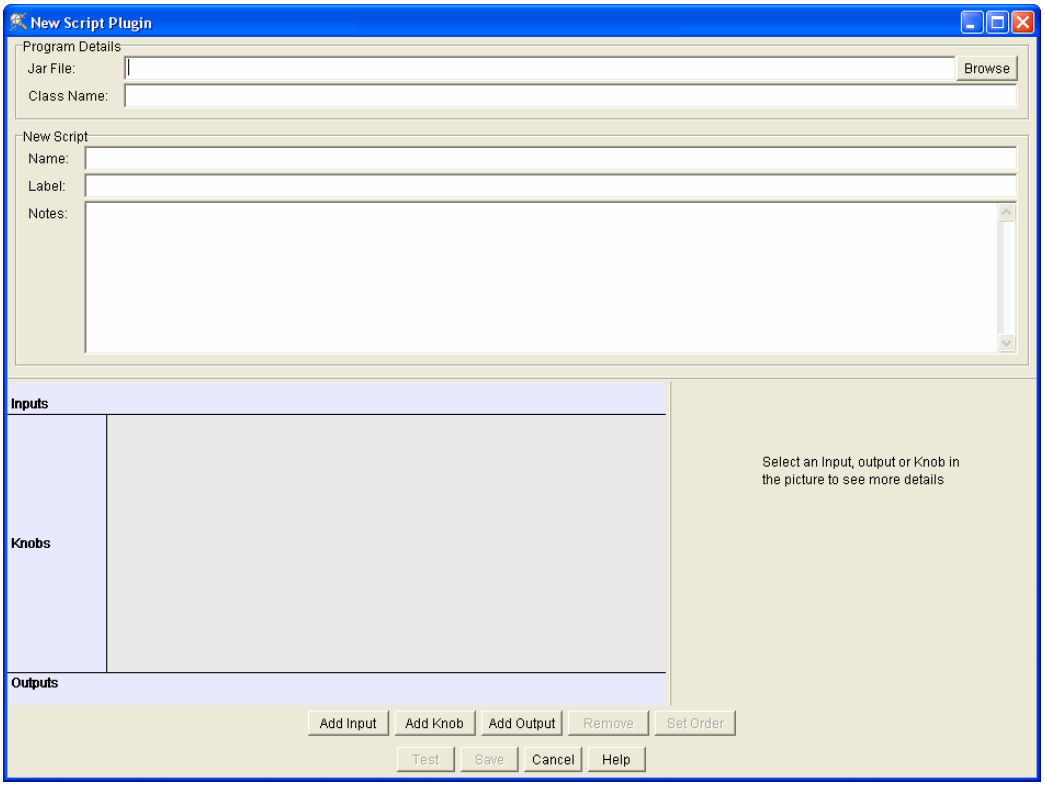


Figure 16-18 New Script Plugin Window

This window contains the following elements:

New Script box

- **Jar File**—Displays the name of the jar file containing the script plugin, including the full path.
- **Browse**—Displays the Browse window, allowing you to choose the correct jar file and have GeneSpring fill in the associated path.

- **Class Name**—Displays the fully qualified class name.
- **Script Name**—Displays the name for the script primitive.
- **Script Label**—Displays the short version of the script name.
- **Script Notes**—Displays the notes associated with this script primitive.

Script Image

The image contains the **Inputs**, **Knobs** and **Outputs** as they are defined for this primitive. This includes showing the correct labels. If an **Input**, **Knob** or **Output** type is undefined the icon will be a question mark, and the **Test** and **Save** buttons will be disabled. Multiple question marks can show at once.

The image is dynamically updated when the **Name** or **Type** of an **Input Socket**, **Output Socket** or **Knob** changes. The Selection is shown the same way as in the ScriptEditor – four dots surrounding the associated triangle. Clicking in the central gray area will deselect whatever **Input**, **Output** or **Knob** is currently selected.

The lines defining the knob, input, and output sections can be selected and dragged to adjust the areas that are allocated to the sections.

Right-clicking on the image gives the following options:

- Move to Start
- Move to End
- Remove (enabled in a vicinity of an existing input or output)
- Add Input/Output (enabled in Input or Output area)
- Add Knob (enabled in the Knobs area)
- Toggle Group (enabled in a vicinity of an existing input or output)

These two options are only enabled when you click in the general vicinity of an existing inputs, knobs, or outputs, and there is more then one input to the script. They enable you to re-order the inputs, without opening the **Set Order** window.

Buttons

- **Add Input**—Adds a new input to the script and a new input to the image, with a question mark indicating that the **Input Type** is undefined.
- **Add Knob**—Adds a new knob to the script and a new knob to the image, with a question mark indicating that the **Knob Type** is undefined.
- **Add Output**—Adds a new output to the script and a new output to the image, with a question mark indicating that the **Output Type** is undefined.
- **Set Order**—Changes the order of the **inputs, knobs, and outputs** associated with this script.
- **Test**—Displays the **Run Script** window associated with the new (and as yet, unsaved) plugin and allows a trial run of the plugin. It is only enabled if all inputs, outputs and knob types are defined, and the **Jar File** and **Class Name** fields are filled in.
- **Save—**Saves the new script plugin.

- **Cancel**—Closes this window without saving the primitive/script plugin.
- **Help—Displays Help** for this window.

Creating Script Plugins

To create a script plugin:

1. Select **File > New Program > New Script Plugin**.

   The New Script Plugin window opens.

2. In the **Program Details** box, do the following:

   a. In the **Jar File** field, click **Browse**, and select the full path name to the jar file that contains the external program you want to run.

   b. In the **Class Name** field, enter the fully-qualified class name (which includes the dots).

3. In the **New Script** box, do the following:

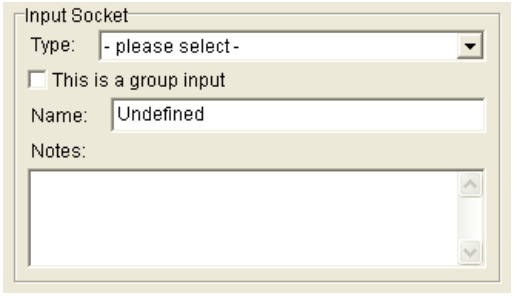
   a. In the **Name** field, enter the name of the plugin.

   b. In the **Label** field, enter the short version of the script name that will be used in the Script Editor graphic interface when this primitive is included within another script.

   c. In the **Notes** field, enter comments about the script.

4. Click the **Add Input** button.

   The Input box appears in the right panel.



5. Do the following:

   a. In the **Input Socket** list, select the type of input you want. The options are:

      - Gene List
      - Gene
      - Condition
      - Experiment
      - Gene Classification
      - Number
      - Boolean
      - Sample

   b. (Optional) If this is a group input, select the **This is a group input** option.

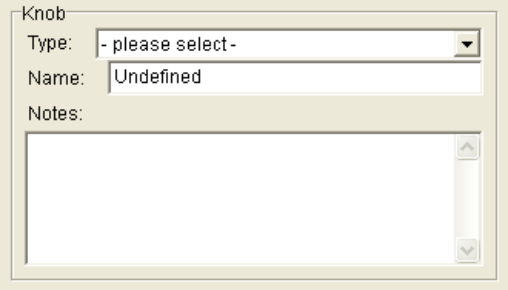
This option enables you to use arrays of objects that are of the same type.

c. In the **Name** field, enter a name for the input.

d. In the **Notes** field, enter comments about the input.

6. Click the **Add Knob** button.

   The Knob section appears in the right panel.



7. Do the following:

   a. In the **Type** list, select the type of knob you want. The options are:

      - Integer
      - Positive Integer
      - Number
      - Positive Number
      - Yes or No
      - Percentage
      - String
      - Enumeration

   b. If you select the **Enumeration** type, then this field is a menu.

      - This menu lists all of the options listed in the **Enumeration Value** table plus a **No Default** option. Each row contains text that you can edit.
      - Clicking on **Add** adds a new row at the bottom of the table, and places the cursor in the new row.
      - Clicking **Remove** removes the selected row.
      - Clicking **Return** while editing one of the rows will automatically add a new row directly below the one you are editing.

   c. If you selected the **Yes or No** type, then this field is a menu with the following three options:

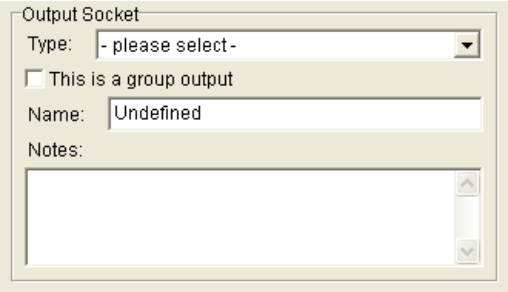
      - No Default
      - Yes
      - No

   d. In the **Name** field, enter a name for the output.

   e. In the **Notes** field, enter comments about the output.

8. Click the **Add Output** button.

   The Output box appears in the right panel.



9. Do the following:

   a. In the **Output Socket** list, select the type of output you want. The options are:

      - Gene
      - Gene list
      - Sample
      - Condition
      - Experiment
      - Gene Classification
      - Number
      - Boolean

   b. (Optional) If this is a group output, select the **This is a group output** option.

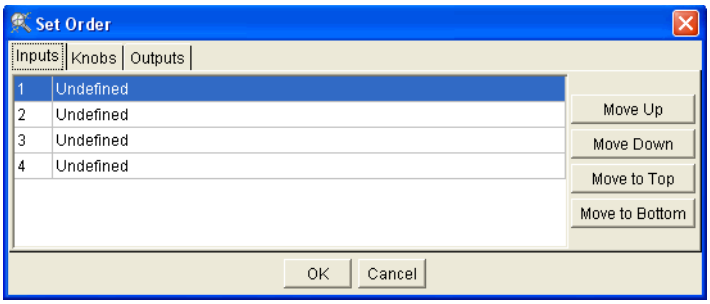
      This option enables the script to output more than one instance of the output type.

   c. In the **Name** field, enter a name for the output.

   d. In the **Notes** field, enter comments about the output.

10. Click the **Set Order** button.

    The Set Order window opens.



11. To set the order of the inputs, do the following:

    a. Click the **Inputs** tab.

    b. Select the row you want to reorder.

c. Click **Move Up**, **Move Down**, **Move to Top**, or **Move to Bottom** to reorder the inputs.

d. Click the **OK** button when you are done.

12. To set the order of the knobs, do the following:

a. Click the **Knobs** tab.

b. Select the row you want to reorder.

c. Click **Move Up**, **Move Down**, **Move to Top**, or **Move to Bottom** to reorder the knobs.

d. Click the **OK** button when you are done.

13. To set the order of the outputs, do the following:

a. Click the **Outputs** tab.

b. Select the row you want to reorder.

c. Click **Move Up**, **Move Down**, **Move to Top**, or **Move to Bottom** to reorder the outputs.

d. Click the **OK** button when you are done.

14. To test the plugin before you save it, click the **Test** button.

15. Click the **Save** button.

The Save Script window opens.

16. In the **Folder** field, enter the name of the folder where you want to store this script plugin.

This step creates a new folder in the Navigator with the name you specify.

Editing Script Plugins

To edit a script plugin:

1. In the **Navigator**, select the folder where your script plugins are saved.
2. Right-click the script plugin you want and select the **Inspect** command.

The External Program Inspector window opens.

3. Edit the **Program Name**, **Author**, and **Research Group** fields.
4. Click **Edit Program Details**.

The Edit Script Plugin window opens.

This window is identical to the New Script Plugin window. Go to "New Script Plugin Window" on page 16-66 for a description of the window elements.

5. Click **Save** when you are done editing the plugin.

17

Exporting Data Files

This chapter explains the different options that are available for exporting data files. It covers the following topics:

- Exporting Images
- Exporting Gene Lists
- Exporting Annotated Gene Lists
- Exporting MAGE-ML Data

Exporting Images

Images within GeneSpring can easily be exported in popular file formats that are readable by a number of applications. Virtually any graphical image can be exported in either pict or png format to enhance the quality of articles for publication. GeneSpring can export expression data in MAGE-ML (Microarray Gene Expression – Markup Language) to facilitate exchange of data with other databases and applications.

You can save Signet Viewer images and import them into a graphics or other program, where you can prepare them for publication. Signet Viewer saves images of pathways, Venn diagrams, the genome browser, and the Colorbar as pct and png files, which can be imported into Microsoft PowerPoint, Word, Publisher, Excel, CorelDRAW, and Adobe Illustrator among other programs.

Saving a Genome Browser Image

You can save a GeneSpring image and import it into a graphics or other program, where you can polish and format it for publication. GeneSpring saves images of pathways, Venn diagrams, the Genome Browser, and the Colorbar as *pct* and *png* files, which can be imported into Microsoft PowerPoint, Word, Publisher, Excel, CorelDRAW, and Adobe Illustrator among other programs.

For example, you can export high resolution images of GeneSpring results, such as hierarchical trees or pathway diagrams in a rasterized format. GeneSpring allows you to export low, medium, or high resolution PNG images to suit your particular needs.

To save a Genome Browser image.

1. Display the image to save in the Genome Browser.
2. Select **File > Save Image** and choose **Browser**.

   The Export Options window opens.
3. Choose an output format for the image.

   The available formats are pict or png. Pict is a vector based graphic format. png is a bitmap based format.

   There are size limits to both formats. Pict files cannot be larger than 450 x 450 inches. png files have a soft limit based on available memory. If the estimated amount of memory to produce the png file is greater than the memory use setting in your Preferences file, a warning dialog opens. In this case, you can attempt to save the png anyway, but if there is not sufficient memory, nothing is saved.
4. Choose an image size from the **Page Size** menu. You have the following options:
   - **Scale to Fit**—calculates the best page size in order to display the graphic and all specified labels. In some cases, this option will specify a page size larger than the maximum. In this case, you must choose another option.
   - **Original Image Size—**lets you save the image exactly as it appears in the Genome Browser.
   - **Original Aspect Ratio—**lets you change the image size, but maintain the original width-to-height ratio displayed in the Genome Browser.
   - **US Letter—**8.5 by 11 inches.
   - **US Legal—**8.5 x 14 inches
   - **A4—**8.3 x 11.7 inches
   - **3 Foot by 5 Foot Poster—**3 ft. by 5 ft.
   - **Custom—**lets you save to any size up to 450 inches by 450 inches.
5. Choose a margin size.

   If you choose **Custom**, enter the appropriate percentage in the **Enter Percentage** box.
6. Choose a landscape or portrait page orientation.
7. Specify whether to show labels, and if so, which labels. The options are:
   - Use Rotated Text for Vertical Labels
   - Force all Text to Show

   You can also specify the text size and font for the labels.
8. Specify the color scheme to use. You can choose either your current color scheme or any of the presets in the menu.
9. Click the **Save** button.

   The Save As window opens.
10. Choose a directory, enter a name for the file, and click the **Save** button.

You may need to save your file as a large custom size, such as 150 x 150 inches, to ensure all data are included in the saved image. Images are saved as vector graphics, which are expandable. Data that are too small to view in the Genome Browser are saved in most cases, and reappear when you expand the image.

Note: Images containing a very large number of genes can require an exceptional amount of memory. The fewer genes included in an image, the smaller the image file.

Saving a Colorbar Image

To save a Colorbar image:

1. Select **File > Save Image** >**Colorbar**.

   The Save As window opens.

2. Choose a directory, enter a name for the file, and click the **Save** button.

Saving a Venn Diagram Image

To save a Venn diagram image:

1. Display the Venn diagram you want to save.
2. Select **File > Save Image** >**Venn Diagram**.

   The Save As window opens.

3. Choose a directory, enter a name for the file, and click the **Save** button.

Saving the Active Window Image

To save the active window image on a Windows system:

1. Press the **Alt** and **Print Window** keys simultaneously to copy a picture of the current active window.
2. Paste the image into any program that accepts graphics and save it.

To save the active window image on a Macintosh system:

1. Press ⌘-**Shift-4-Caps Lock** simultaneously.

   The cursor changes to a bull’s-eye.

2. Click on a GeneSpring window to save the image as a file on your hard drive called “Picture”.

   You must rename this file, otherwise it is overwritten each time you repeat this procedure.

Saving the Entire Window Image

To save the entire window image on a Windows system:

1. Press the **Print Window** key to copy a picture of the current active window.
2. Paste the image into any program that accepts graphics and save it.

To save the entire window image on a Macintosh system:

1. Press ⌘-**Shift-3-Caps Lock** simultaneously.

   The cursor changes to a bull's-eye.

2. Click on a GeneSpring window to save the image as a file on your hard drive called "Picture".

   You must rename this file, otherwise it is overwritten each time you repeat this procedure.

Printing Images

You can print an image of the Genome Browser, the Genome Browser with the Colorbar, or the display window. Such images can be useful for reports or handouts. Use a high-resolution color printer to print GeneSpring images.

You can print an image of the genome browser, the genome browser with the Colorbar, or the display window. Such images can be useful for reports or handouts. Use a high-resolution color printer to print Signet Viewer images.

Printing the Genome Browser and Colorbar

To print the Genome Browser and Colorbar:

1. Do one of the following:
   - Select the **File > Print Image > Browser**
   - Select the **File > Print Image > Colorbar**
   - Select the **File > Print Image > Browser and Colorbar**.

   The Print window opens.

2. Select a printer and click the **OK** button.

Printing the Active Window Image

To print the active window image on a Windows system:

1. Press the **Alt** and **Print Window** keys simultaneously to copy a picture of the current active window.
2. Paste the image into any program that accepts graphics and save it.
3. Print the image using your local or network printer.

To print the active window image on a Macintosh system:

1. Press ⌘-**Shift-4-Caps Lock** simultaneously.

   The cursor changes to a bull's-eye.

2. Click on a GeneSpring window to save the image as a file on your hard drive called "Picture".

   You must rename this file, otherwise it is overwritten each time you repeat this procedure.

3. Print the image using your local or network printer.

Printing the Entire Window Image

To print the entire window image on a Windows system:

1. Press the **Print Window** key to copy a picture of the current active window.
2. Paste the image into any program that accepts graphics and save it.
3. Print the image using your local or network printer.

To print the entire window image on a Macintosh system:

1. Press ⌘-**Shift-3-Caps Lock** simultaneously.

   The cursor changes to a bull's-eye.

2. Click on a GeneSpring window to save the image as a file on your hard drive called "Picture".

   You must rename this file, otherwise it is overwritten each time you repeat this procedure.

3. Print the image using your local or network printer.

Exporting Gene Lists

You can make gene lists and annotated gene lists available to another application. An annotated list includes functional descriptions, as well as standard deviation, standard error and other information associated with the gene list (when available).

Dragging Gene Lists out of GeneSpring

You can drag a gene list out of the Navigator to your desktop or use the Export to Zip short cut menu command. If dragged to the desktop, the gene list is saved as a zip file. The resulting list contains only the gene identifiers and associated values.

Dragging and dropping a gene list does not produce the same list as does the Copy Annotated Gene List function.

Copying and Pasting Gene Lists

To copy a gene list from the Navigator:

1. In the **Navigator**, open the **Gene Lists** folder.
2. Select the gene list you want.
3. Select **Edit > Copy > Copy Gene List**.
4. Paste the list into a new application.

Exporting Annotated Gene Lists

GeneSpring enables you to export annotated gene lists to the clipboard or to a file. Annotated gene lists can contain expression data, along with annotation data. You can pick and choose exactly what data you want to export. This section explains the different options that are available.

Copy Annotated Gene List Window

The Copy Annotated Gene List window (Figure 17-1) lets you copy and save information from an annotated gene list. The type and amount of information that appears in this window varies depending on your genome and the way that genome was loaded into GeneSpring.

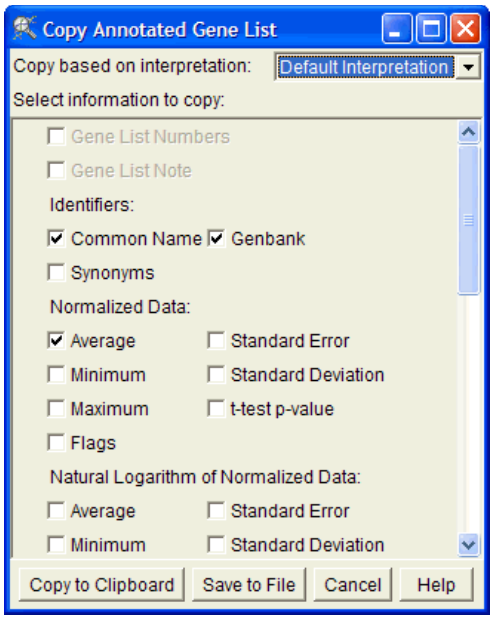


Figure 17-1 Copy Annotated Gene List Window

Your options for copying and saving information with an annotated gene list are listed in the Copy Annotated Gene List window. Descriptions of these items can be found by clicking **Help**. The type and amount of information listed varies depending on your

genome and the way that genome was loaded into Signet. The Systematic Name is always saved in the first column of a gene list.

The following section describes the elements in this window.

Copy Based on Interpretation

If you choose to include expression values, this section determines from which interpretation the expression values come. If you do not need to copy the expression values and are only interested in the annotation values, you do not need to set the correct interpretation.

- **Default Interpretation**—The Default Interpretation is the first item listed under the experiment in the Navigator. It may be most convenient to set up your most frequently used interpretation as your Default Interpretation. You can rename the Default Interpretation, but you cannot delete it.
- **All Samples**—This interpretation makes all parameters non-continuous, so that each parameter is viewed and analyzed individually. The All Samples interpretation cannot be changed, renamed, or deleted.

General

- **Gene List Numbers**—The values (if any) that GeneSpring has associated with this gene list. This check box is active only if you have associated values.

  Some examples of associated numbers are correlation coefficients, p-values, fold change ratios, or in the case of a regulatory sequence search, the number of base pairs before the promoter region. Associated numbers can be found by double-clicking a gene list to bring up the Gene List Inspector.
- **Gene List Note**—Any notes attached to a gene list.

Identifiers

- **Systematic Name**—The unique identifier for the gene in this genome or array. It is recommend that the gene's systematic name be used to label the gene's expression values in your experiment data files.
- **Common Name**—An alternative way of referring to this gene. Genes are not required to have a common name, and common names do not have to be unique, although duplicated common names may lead to confusion if the common name is how the gene is referred to in the experiment files. Usually, the common name annotation column can be used to store the HUGO gene symbol or some other official gene identifier.
- **GenBank Accession Number**—The GenBank or EMBL identifier for this gene, if known. If the GenBank identifiers for your genes were not used as either their systematic or common names, then including the GenBank Accession Number in this field allows you to update the information about this particular gene directly from GenBank. See "Updating Annotations with GeneSpider" on page 11-4 for more information.

Normalized Data

- **Average—**The mean of any normalized replicates in the selected experiment interpretation.
- **Minimum—**The minimum normalized signal values for each gene.
- **Maximum—**The maximum normalized signal values for each gene.
- **Flags**—Flags associated with each gene
- **Standard Error—**The standard error of the normalized values for each gene in the selected interpretation.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the raw data values for each gene in the selected interpretation.
- **t-test p-value**—The statistical test of differential expression for a specific condition.

Natural Logarithm of Normalized Data

Choosing these columns will include the (natural) log values of the normalized data.

- **Average—**The mean of any normalized replicates in the experiment.
- **Minimum—**The minimum normalized signal values for each gene.
- **Maximum—**The maximum normalized signal values for each gene.
- **Standard Error—**The standard error of the normalized values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the normalized values for each gene.

Raw Data

- **Average—**The mean of any raw data replicates in the experiment.
- **Minimum—**The minimum raw data signal values for each gene.
- **Maximum—**The maximum raw data signal values for each gene.
- **Standard Error—**The standard error of the raw data values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the raw data values for each gene.

Control Value

- **Average—**The mean of any control value replicates in the experiment.
- **Minimum—**The minimum control value signal values for each gene.
- **Maximum—**The maximum control value signal values for each gene.
- **Standard Error—**The standard error of the control values for each gene.
- **Standard Deviation—**The standard deviation (the square root of the variance) of the control values for each gene.

Annotations

The columns of annotation that can be used depend on which annotation columns are present in your genome. The Annotations sections list all the annotation columns that are defined for your genome, regardless of whether they contain data for the selected genes.

Some of the standard annotation columns that are always present are:

- **Map Position—**A gene's mapping information.
- **Chromosome—**The chromosome on which a gene is located, if known.
- **User Notes**—Any additional notes you may have associated with a gene.
- **EC—**A gene's EC (Enzyme Commission) number, if known.
- **Description—**A gene's description, if known.
- **Product—**The protein product coded for by a gene, if known.
- **Phenotype—**A description of a gene's phenotype, if known.
- **Function—**A description of the function of a gene's product, if known.
- **Keywords—**Keywords associated with a gene, if known.
- **PubMed ID—**A gene's PubMed identifier.
- **Custom Field 1, Custom Field 2, Custom Field 3—**Any information you choose to place here for your own use.
- **Type**—The feature type from the GenBank file.
- **DB id—**A reference used to identify a gene within Signet.
- **GO Biological Process**—The Gene Ontology Biological Process classification
- **GO Molecular Function**—The Gene Ontology Molecular Function classification
- **GO Cellular Component**—The Gene Ontology Cellular Component classification
- **RefSeq**—The gene's NCBI Reference Sequence project identifier.
- **UniGene**—The gene's UniGene cluster identifier.

Copying Annotated Gene Lists

To copy an annotated gene list:

1. In the **Navigator**, open the **Gene Lists** folder.
2. Select the annotated gene list you want to copy.
3. Select **Edit > Copy > Copy Annotated Gene List**.

   The Copy Annotated Gene List window opens.
4. In the **Copy based on interpretation** menu, select the interpretation you want to use.

   Go to "Setting Up Experiment Interpretations" on page 5-65 for information on experiment interpretations.
5. In the **Select Information to copy** box, select the options you want.

Go to "Copy Annotated Gene List Window" on page 17-6 for a description of the options.

6. Click **Copy to Clipboard**.
7. Paste the list into another application.

   The resulting text file can be opened in any program that accepts tab-delimited text, such as a spreadsheet or a word processing program.

Saving Annotated Gene Lists

To save an annotated gene list:

1. In the **Navigator**, open the **Gene Lists** folder.
2. Select the annotated gene list you want to copy.
3. Select **Edit > Copy > Copy Annotated Gene List**.

   The Copy Annotated Gene List window opens.
4. In the **Copy based on interpretation** menu, select the interpretation you want to use.

   Go to "Setting Up Experiment Interpretations" on page 5-65 for information on experiment interpretations.
5. In the **Select Information to copy** box, select the options you want.

   Go to "Copy Annotated Gene List Window" on page 17-6 for a description of the options.
6. Click the **Save to File** button.

   The Save Annotated Gene List window open.
7. Choose a directory, enter a name for the file, and click the **Save** button.

The resulting text file can be opened in any program that accepts tab-delimited text, such as spreadsheet and word processing programs.

Exporting MAGE-ML Data

MAGE-ML (Microarray Gene Expression Markup Language) is a markup language based on XML and designed to describe and communicate information about microarray experiments. Using MAGE-ML, your experimental data can include information about microarray designs, manufacturing information, and experiment setup and execution information as well as gene expression data and analysis results. Certain publications and laboratories require results to be published to Array Express or Gene Expression Omnibus in MAGE-ML format.

Array Express is a new public repository for microarray based gene expression data. Incyte provided funding for its creation, and it is now funded by EMBL and EBI. Array Express accepts only data submitted via MIAMExpress (a web-based submission interface) or via FTP in MAGE-ML format.

MAGE-ML is extremely broad in its definitions. As a result, EBI/Array Express has defined its own version of MAGE-ML. Array Express will only accept submissions that are both MIAME compliant and conform to their MAGE-ML standard.

GeneSpring currently supports export only of MAGE-ML data. It does not fully support export of data in a format that is applicable for ArrayExpress submission.

MAGE-ML Data for Publication

When you a export MAGE-ML file from GeneSpring, the following files are included:

- An XML file in MAGE-ML format describing the experiment
- A supporting tab-delimited text file containing the actual expression data
- All raw data files that were originally imported into GeneSpring to create the samples in the experiment

MAGE-ML files may be submitted to EBI at http://www.ebi.ac.uk/miamexpress/cgi-bin/mx.cgi.

CompositeSequence Identifiers

By default, GeneSpring generates a MAGE-ML CompositeSequence identifier of the form, *CS:01:systematicname*, for each gene. As with all identifiers, if a string is entered into the Identifier prefix" field, it will be prepended with a colon character to each identifier generated this way.

In general, CompositeSequence identifiers generated this way will not map to externally resolvable probesets. The Export window now allows identifiers for each gene to be specified by an annotation column from the Master Table of Genes.

To use CompositeSequence identifiers, the gene annotations must contain a column with the CompositeSequence Identifiers. Go to Chapter 11, "Annotating Genes" for information on how to edit the gene annotations.

To use an annotation column as the source for your CompositeSequence identifiers, check the Use annotation column check box in the Export window and select the appropriate field to use from the pull-down list. The values in the selected column cannot contain duplicates, and the individual annotation values cannot contain spaces.

If values are missing for some genes, those genes will not appear in the output. CompositeSequence identifiers generated in this fashion will not have any prefix appended to them if one is specified in the Identifier Prefix field.

Affymetrix Probesets

For Affymetrix probesets, recommended values for submission to ArrayExpress have the form:

`Affymetrix:CompositeSequence:design:probesetname`

Where *probesetname* is the Affymetrix probeset identifier (usually the systematic name) and *design* is one of the following:

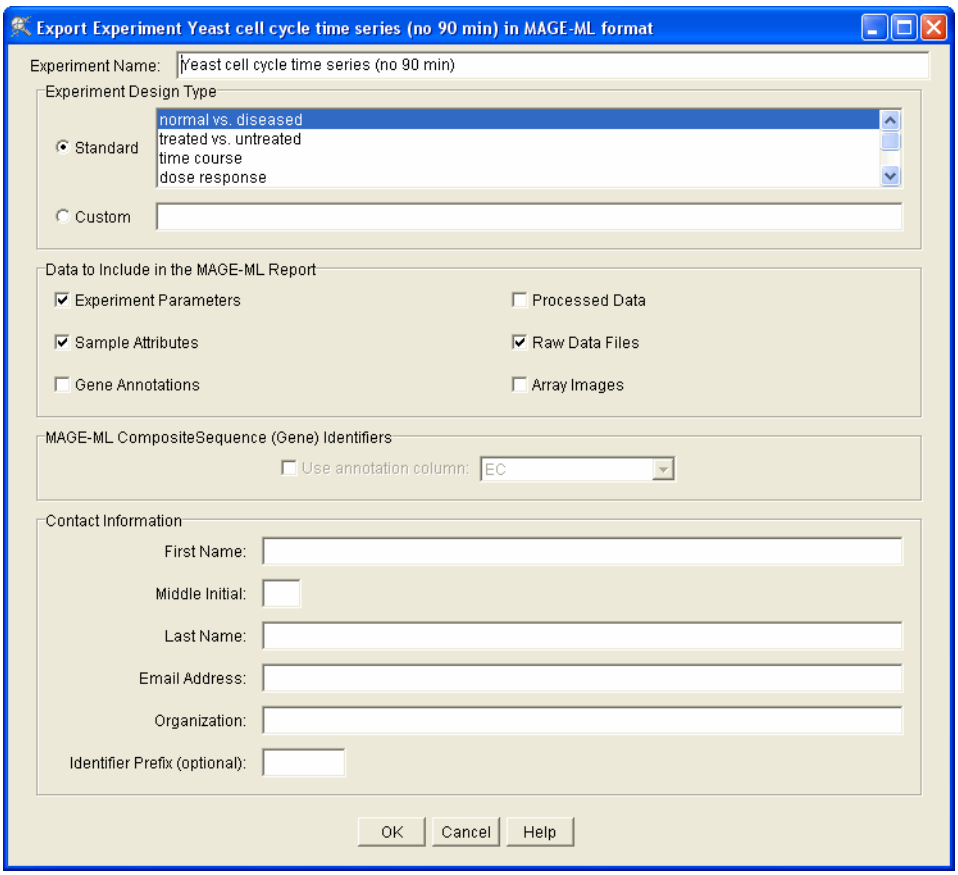
HG_U95A	Affymetrix Genechip® Human Genome U95A
HG_U95B	Affymetrix Genechip® Human Genome U95B
HG_U95C	Affymetrix Genechip® Human Genome U95C
HG_U95D	Affymetrix Genechip® Human Genome U95D
HG_U95E	Affymetrix Genechip® Human Genome U95E
HG_U95Av2	Affymetrix Genechip® Human Genome U95Av2
HuGeneFL	Affymetrix Genechip® HuGeneFL
HG-U133A	Affymetrix GeneChip® Human Genome HG-U133A
HG-U133B	Affymetrix GeneChip® Human Genome HG-U133B
MG_U74A	Affymetrix Genechip® Murine Genome U74A
MG_U74B	Affymetrix Genechip® Murine Genome U74B
MG_U74C	Affymetrix Genechip® Murine Genome U74C
MG_U74Av2	Affymetrix Genechip® Murine Genome U74Av2
MG_U74Bv2	Affymetrix Genechip® Murine Genome U74Bv2
MG_U74Cv2	Affymetrix Genechip® Murine Genome U74Cv2
MOE430A	Affymetrix GeneChip® Mouse Expression Array MOE430A
MOE430B	Affymetrix GeneChip® Mouse Expression Array MOE430B
RG_U34A	Affymetrix Genechip® Rat Genome U34A
RG_U34B	Affymetrix Genechip® Rat Genome U34B
RG_U34C	Affymetrix Genechip® Rat Genome U34C
RT_U34	Affymetrix Genechip® Rat Toxicology U34
RN_U34	Affymetrix GeneChip® Rat Neurobiology U34
RAE230A	Affymetrix GeneChip® Mouse Expression Array RAE230A
RAE230B	Affymetrix GeneChip® Mouse Expression Array RAE230B
ATH1-121501	Affymetrix Genechip® Arabidopsis Genome [ATH1]
AG	Affymetrix Genechip® Arabidopsis Genome [AG1]
DrosGenome1	Affymetrix Genechip® Drosophila Genome
YG_S98	Affymetrix Genechip® Yeast Genome S98
Pae_G1a	Affymetrix GeneChip® P.Aeruginosa Genome

Exporting MAGE-ML Data

To export MAGE-ML data:

1. In the **Navigator**, open the **Experiments** folder
2. Right-click an experiment and select **Export as MAGE-ML**.

   The MAGE-ML Export window opens.



3. Enter an experiment name in the field provided.
4. For submission of new data or updated information to the European Bioinformatics Institute (EBI)[1], you must complete steps 5 through step 8.
5. In the **Experiment Design Type box, do the following:**
   - Select the **Standard** option and choose a design from the list.
   - Select the **Custom** option and enter a description.

     Note: This step is also required for MIAME compliance.

1. EBI is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL). The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures.

6. Check *only* the **Experimental Parameters**, **Sample Attributes**, and **Raw Data Files** boxes.
7. To map the gene identifiers in the experiment MAGE-ML file with the identifiers used by EBI do the following
   a. In the **MAGE-ML CompositeSequence (Gene) Identifiers** option. select **Use annotation column**.

      Note: This option is only enabled if you choose to output the gene annotations.

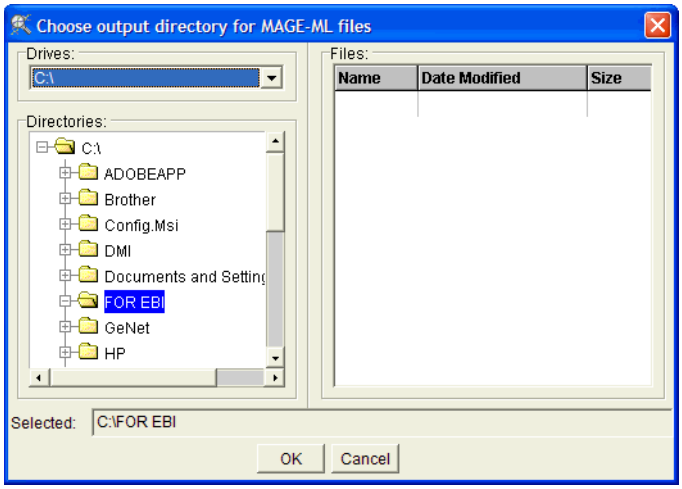
   b. Select which column of the **Master Table of Genes** to use as the gene identifier in the MAGE-ML file.

      Note: It is recommended that you designate a special column in the annotations file for EBI CompositSequence names. Go to "Managing Genes and Annotations" on page 6-35 for more information.
8. Fill in all of the **Contact Information**, including your lab's EBI accession number. If you do not specify an accession number, your experiment's accession number is used by default.

   Note: If you do not have an EBI accession number, you must contact EBI to obtain one.

   By default, GeneSpring's MAGE-ML export feature specifies the appropriate options to produce EBI-compliant MAGE-ML data. You have the option of including more information than EBI requires, but if you do so, your data will not be accepted by Array Express.
9. Click **OK**.

   The Choose Output Directory window opens.



10. Select the directory in which the MAGE-ML export files will be saved.

An EBI-compliant MAGE-ML experiment must contain at least the following three files:

- An XML file describing the experiment
- An MAGE-ML.dtd file, which defines the MAGE-ML format
- A plain text file containing the raw experimental data

Your export may also include raw data files and/or sample or array images.

Note: Make sure the Systematic names of your genes are the same identifiers used by the chip manufacturer. This eases mapping from the vendor MAGE-ML file to the experiment MAGE-ML file.

11. Click the **OK** button.

The MAGE-ML files are saved to the directory you specified.

18

Using Signet

This chapter explains how to manage data when using Signet. It covers the following topics:

- Signet
- Configuring GeneSpring to Access Signet
- Logging in to Signet
- Uploading Data to Signet
- Bulk Upload of Data to Signet
- Editing Data in Signet
- Changing Ownership of Data in Signet

Signet

Signet is a scalable workspace for GeneSpring users that streamlines microarray research at large or multi-campus organizations. It provides facilities for robust data analysis, collaborative workflow management, automated research procedures, and secure data administration. Signet scales to meet the demands of both high-throughput sample volumes and increasing numbers of users. You can publish any data object from GeneSpring to Signet.

Working with Signet is easier than ever. Now you can synchronize your local data with data on Signet using just one click. At the end of an analysis, a single click allows you to upload all new data to Signet, assuring that your data and analysis are safely stored on Signet and immediately available for your collaborators.

Improved Search Signet functionality allows you to find any gene list, sample, experiment or data object on Signet no matter which genome it is associated with. Now, you can also use a genome on Signet as your default genome—simply log in to Signet as GeneSpring starts up.

Configuring GeneSpring to Access Signet

The Preferences window in GeneSpring provides a tab for specifying Signet preferences. In the Signet tab (Figure 18-1), you can specify the default Signet server to connect to, as well as the addresses of other Signet servers to which you want access.

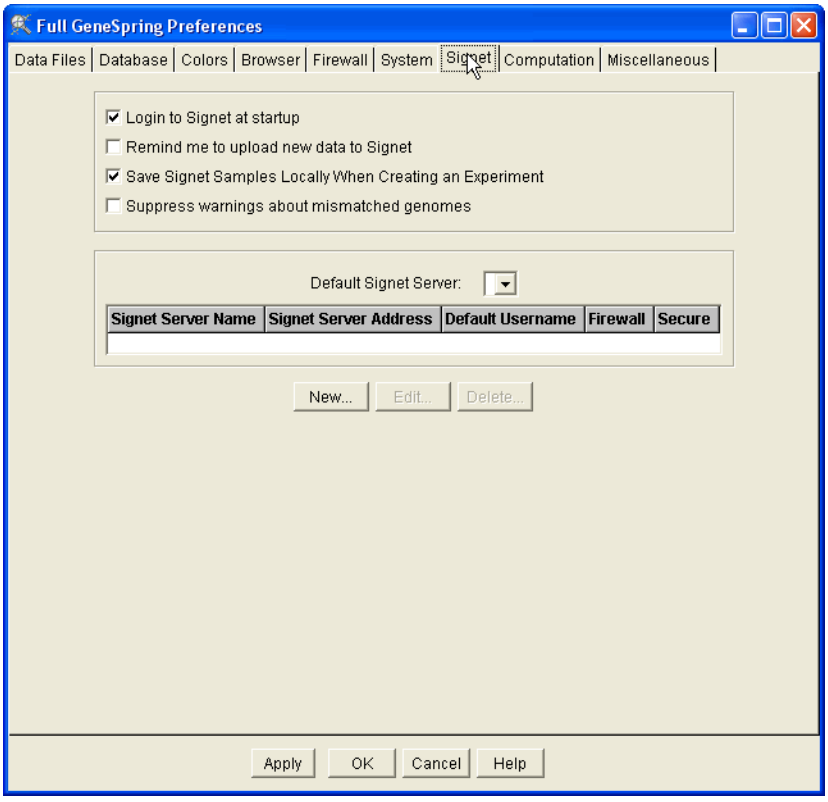


Figure 18-1 Preferences Window: Signet Tab

The Signet tab provides the following options:

- To enable GeneSpring to automatically connect to the default Signet server at startup, check the **Login to Signet at Startup** box. Select the default Signet server from the menu.
- To enable GeneSpring to invoke the Bulk Upload to Signet window when you quit GeneSpring, check the **Remind me to upload new data to Signet** box.
- If your local copy of a genome contains a different list of genes than the version of that genome on Signet, GeneSpring will warn you every time you open the genome. To disable these warnings, select the **Suppress warnings about mismatched genomes** check box.
- When you create a new experiment in GeneSpring using samples from a Signet server, GeneSpring saves local copies of those samples by default. This can cause slow performance when saving large experiments. To disable this feature, clear the **Save Signet Samples Locally When Creating an Experiment** box.

Adding a New Signet Server

To add a new Signet server:

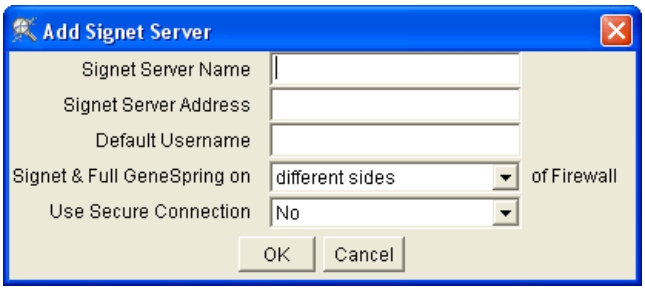
1. Select **Edit > Preferences**.

   The Preferences window opens.

2. Select the **Signet** tab.

3. Click the **New** button.

   The Add Signet Server window opens.



4. Enter the following information:

- **Signet Server Name**—Specifies the name of the server to connect to
- **Signet Server Address**—Specifies the IP address or host name of the Signet server. For example: 127.0.0.1. Do not enter “http://” before this address. Contact your system administrator to obtain this information.
- **Default User Name**—Specifies the user name to connect to Signet.
- **Signet & GeneSpring on**—Specifies whether the Signet server is on the same side of your firewall as GeneSpring or is on different sides.
  - If Signet and GeneSpring are on different sides of the firewall, you need to configure the firewall settings and then select the **different sides** option. For information on configuring firewall settings, go to “Firewall Tab” on page 3-20.
  - If you do not have a firewall, select the **same side** option.
- **Use Secure Connection**—Specifies whether communication between GeneSpring and Signet should be secured or not. If it is secured, communication between Signet and GeneSpring uses HTTPS. This option will only work if Signet has been configured to support secure communication.

5. Click the **OK** button.

   Once the server is defined, you only need to select the server you want and log in to Signet with your user name and password. You will not be required to define the server again.

Editing Settings for a Signet Server

To edit settings for a Signet server:

1. Select **Edit > Preferences**.

   The Preferences window opens.

2. Select the **Signet** tab.
3. Select it from the list and click the **Edit** button.

   The Edit Signet Server window opens.

4. Edit the settings you want.
5. Click the **OK** button when you are done.

Deleting a Signet Server

To delete a Signet server:

1. Select **Edit > Preferences**.

   The Preferences window opens.

2. Select the **Signet** tab.
3. Select it from the list and click the **Delete** button.

Logging in to Signet

To log in to Signet:

1. Do one of the following:
   - Select **File > Login to Signet**.
   - Select **File > Genome Manager**, and click the **Login to Signet** button.

   The Login window opens.

2. From the **Signet Server** list, select the server you want.
3. Enter your **User Name** and **Password**.
4. Click the **OK** button.

Uploading Data to Signet

To upload data to Signet:

1. Log in to **Signet**.

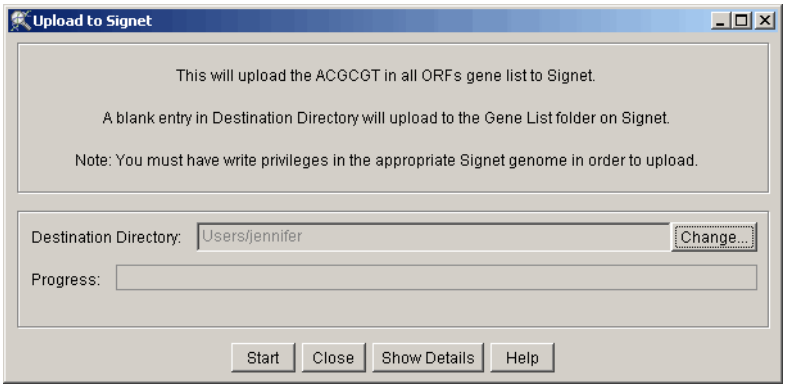
   Go to "Logging in to Signet" on page 18-4 for information.

2. In the **GeneSpring Navigator**, locate the data object you want to upload.

3. Do one of the following:

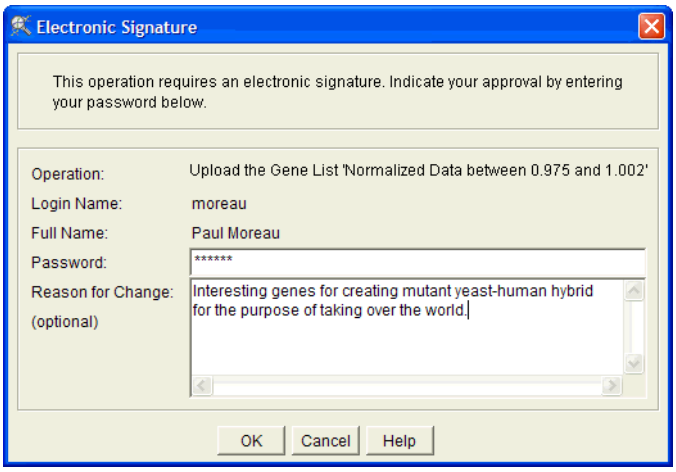
   - Right-click the data object and select **Upload to Signet**.
   - Select the data object you want and drag it into the appropriate Signet folder (displayed in *italics*) in GeneSpring.

     The Signet Upload window opens.


Upload to Signet
This will upload the ACGCGT in all ORFs gene list to Signet.
A blank entry in Destination Directory will upload to the Gene List folder on Signet.
Note: You must have write privileges in the appropriate Signet genome in order to upload.
Destination Directory:
Users/jennifer
Change...
Progress:
Start
Close
Show Details
Help

4. Click the **Start** button.

   The upload process begins. If you are uploading to a regulatory-compliant server, the Electronic Signature window opens.



5. Do the following:

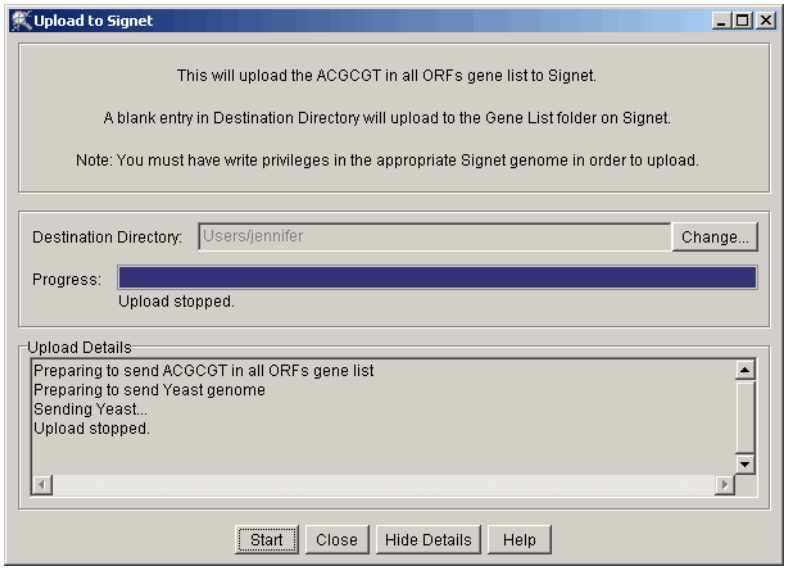
   a. Enter your Signet password.

      Your Signet password serves as your electronic signature.

   b. (Optional) Enter a brief text message describing the purpose of your action.

   For additional information about electronic signatures and other regulatory compliance features in Signet, go to the *Signet Administration Manual*.

   The Progress bar notifies you when the upload is complete.


Upload to Signet
This will upload the ACGCGT in all ORFs gene list to Signet.
A blank entry in Destination Directory will upload to the Gene List folder on Signet.
Note: You must have write privileges in the appropriate Signet genome in order to upload.
Destination Directory:
Users/jennifer
Change...
Progress:
Upload stopped.
Upload Details
Preparing to send ACGCGT in all ORFs gene list
Preparing to send Yeast genome
Sending Yeast...
Upload stopped.
Start
Close
Hide Details
Help

   If you have trouble uploading data, ask your administrator to confirm that you have access to the target directory on Signet.

   Once your data is uploaded to Signet, you can view them in the Navigator of the main GeneSpring window. Folder and file names appear italicized to distinguish them from locally stored folders.

6. If errors occur during the upload process, you can view them by clicking the **Show Details** button.

Bulk Upload of Data to Signet

The Bulk Upload function lets you upload large amounts of data to Signet at once. Administrators can also upload new genomes, and will have additional options in this Bulk Upload to Signet window not available to other users.

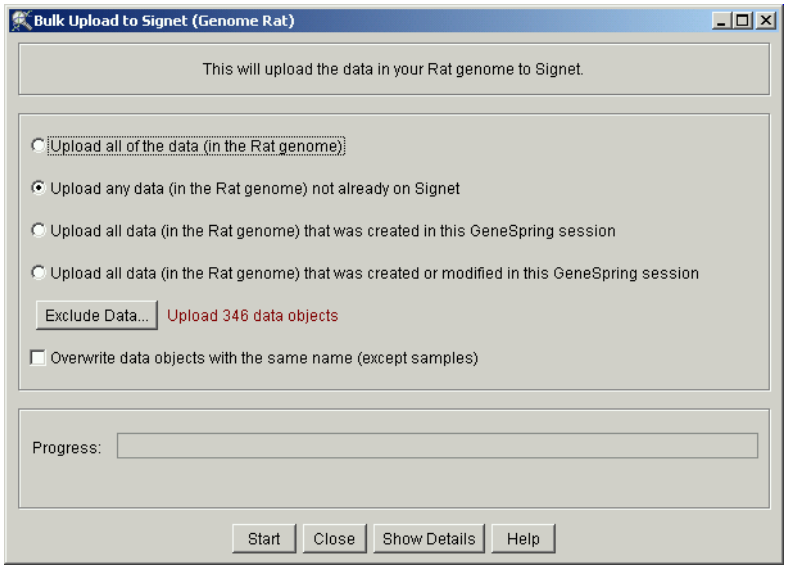
To upload genomes to Signet:

1. Log in to **Signet**.

   Go to "Logging in to Signet" on page 18-4 for information.

2. Select **File > Bulk Upload to Signet**.

   The Bulk Upload to Signet window opens.


Bulk Upload to Signet (Genome Rat)
This will upload the data in your Rat genome to Signet.
Upload all of the data (in the Rat genome)
Upload any data (in the Rat genome) not already on Signet
Upload all data (in the Rat genome) that was created in this GeneSpring session
Upload all data (in the Rat genome) that was created or modified in this GeneSpring session
Exclude Data...
Upload 346 data objects
Overwrite data objects with the same name (except samples)
Progress:
Start
Close
Show Details
Help

3. Do one of the following:

   - To upload only the genome, select the **Upload only the genome** option.
   - To upload the genome and associated data, select the **Upload the genome and all of the data**.

     Data refers to the samples, experiments, gene lists, and other data objects associated with the genome. If you only want to update the gene annotation, do not select this option.

   - To upload the genome and any new data not already in Signet, select the **Upload the genome and all data not already on Signet** option.

     With this option, GeneSpring verifies if the samples, experiments, or other data objects are already stored on Signet. If they are not, GeneSpring will only upload those objects. This option enables you to synchronize the local data with the data on Signet.

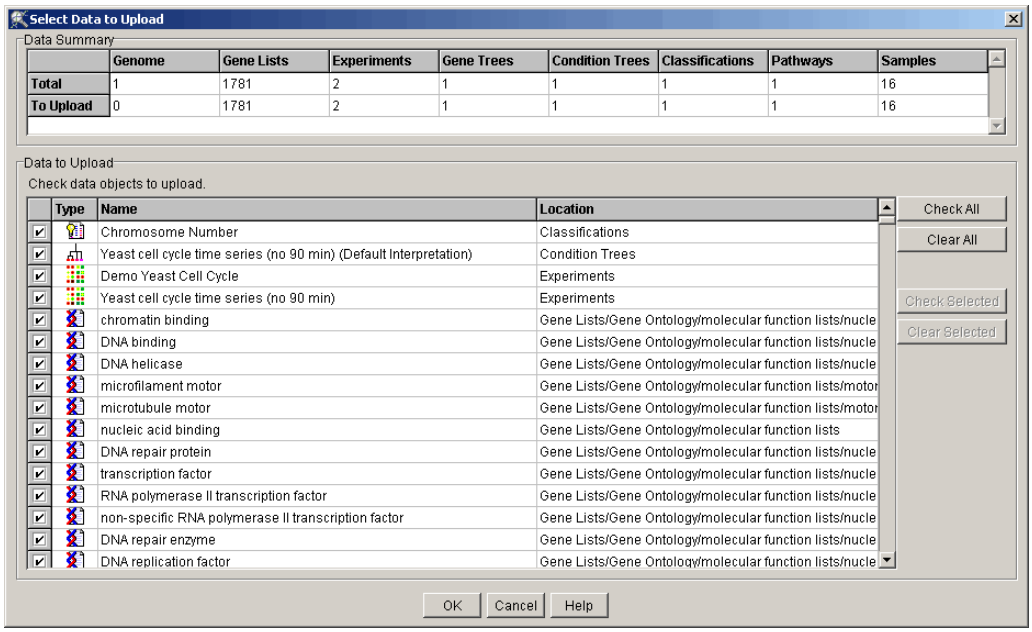
     Note: If you select this option, but do not want to upload certain objects, go to step 4 for instructions on how to exclude data.

- To upload the genome and any new data created during this session, select the **Upload the genome and all data created in this GeneSpring session** option.
- To upload the genome and all data created or modified during this session, select the **Upload the genome and all data that were created or modified in this GeneSpring session** option.

4. To manually select the data objects you do not want to upload, do the following:

   a. Click the **Exclude Data** button.

      The Select Data to Upload window opens.



   b. To include files and data objects to upload, select the check box next to the name.

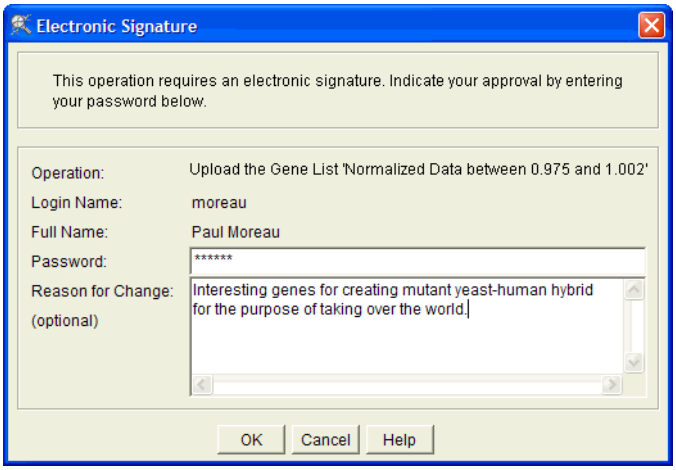
   c. To exclude files or data objects, clear the check box next to the name.

   d. Click the **OK** button.

5. To overwrite any previously uploaded data objects with the same name (other than samples), select the **Overwrite data objects with same name** option.

6. Click the **Start** button.

   The upload process begins. If you are uploading to a regulatory-compliant Signet server, the Electronic Signature window opens.



7. Do the following:

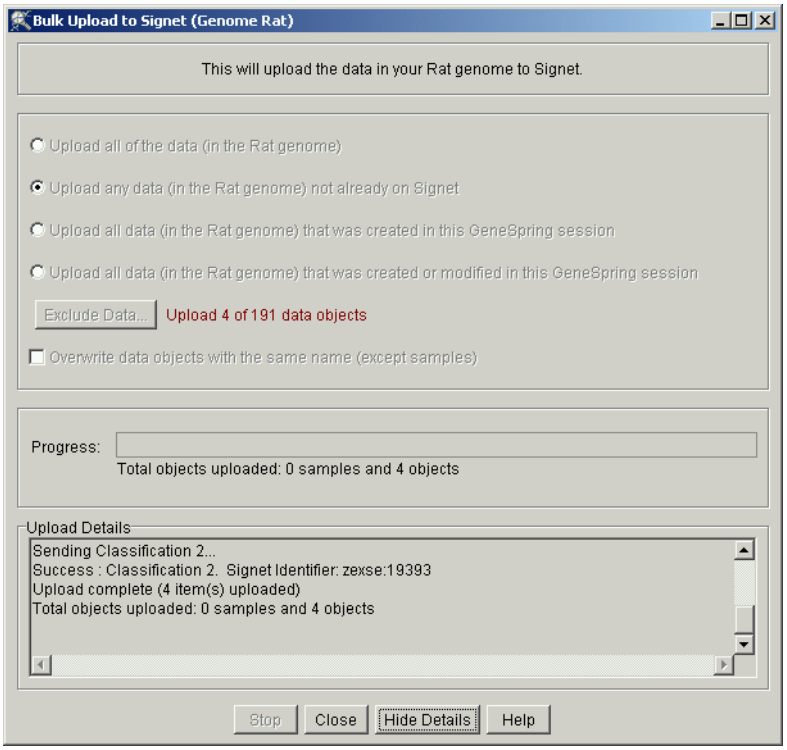
   a. Enter your Signet password.

      Your Signet password serves as your electronic signature.

   b. (Optional) Enter a brief text message describing the purpose of your action.

   For additional information about electronic signatures and other regulatory compliance features in Signet, go to the *Signet Administration Manual*.

   The Progress bar notifies you when the upload is complete.


Bulk Upload to Signet (Genome Rat)
This will upload the data in your Rat genome to Signet.
Upload all of the data (in the Rat genome)
Upload any data (in the Rat genome) not already on Signet
Upload all data (in the Rat genome) that was created in this GeneSpring session
Upload all data (in the Rat genome) that was created or modified in this GeneSpring session
Exclude Data...
Upload 4 of 191 data objects
Overwrite data objects with the same name (except samples)
Progress:
Total objects uploaded: 0 samples and 4 objects
Upload Details
Sending Classification 2...
Success : Classification 2. Signet Identifier: zexse:19393
Upload complete (4 item(s) uploaded)
Total objects uploaded: 0 samples and 4 objects
Stop
Close
Hide Details
Help

If you are having trouble uploading, ask your administrator to confirm that you have access to the target directory on Signet.

Once your data is uploaded to Signet, you can view them in the Navigator of the main GeneSpring window. Folder and object names appear italicized to distinguish them from locally stored folders.

8. If errors occur during the upload process, you can view them by clicking the **Show Details** button.

Editing Data in Signet

Experiments and samples stored on Signet can be modified. However, you can only modify a sample or an experiment if you own it (or are an administrator).

To enable editing:

1. Open a window in which the sample or experiment can be edited, such as the Sample Inspector or Experiment Normalization window.
2. Click the **Edit** button.
   - Only one person can edit an object at one time. If you click the Edit button and someone else is editing it, you will receive an error message.
   - One you begin editing, no one else can edit that experiment or sample until you close all editing windows for it.

Changing Ownership of Data in Signet

If you have data stored on Signet, you can give other users permission to access and change that data in GeneSpring. To change ownership, you must be either an administrator or the owner of the data object.

This section explains how to change ownership using the Navigator in the main GeneSpring window. This same functionality is also available through the Experiment Inspector and the Sample Manager.

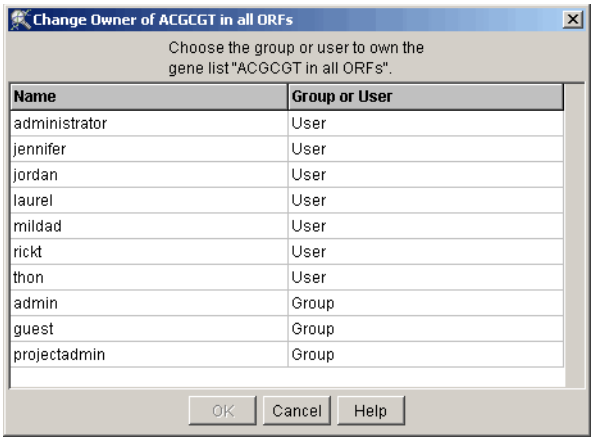
To change ownership of data objects:

1. Log in to **Signet**.

   Go to "Logging in to Signet" on page 18-4 for information.

2. In the **Navigator,** right-click the data object you want, and then select the **Change Owner** command.

   The Change Owner window opens.



3. Choose the group or user to add as an owner.
4. Click the **OK** button.

A

Custom Databases and GeneSpring

You can load experiments into GeneSpring from your company's database. To do this, you must set up a dataloader.xml file prior to starting GeneSpring.

Database

A database is an organized collection of information. Essentially, it is a collection of records. In database terms, a record consists of all the useful information you can gather about a particular item. Each little bit of information making up a record is called a field. An example of a non-computerized database would be your address book. Each record represents one of your contacts, and each record consists of many fields such as name, address, number, and so on.

Computer databases automatically keep records organized and enable you to search for or pull out particular records based on any field in the record. The software allowing you to create and maintain databases is called a *Database Management System*, or DBMS. In database terminology, a *file* is called a *table*. Each *record* in the file is called a *row*, and each *field* is called a *column*.

A relational database is the most common type of database in client/server systems. Simply stated, in this type of database, relationships are established between tables based on common information.

Open Database Connectivity

Open Database Connectivity (ODBC) is an Application Programming Interface (API) allowing a programmer to abstract a program from a database. When writing code to interact with a database, you usually must add code that talks to a particular database using a proprietary language. If you want your program to talk to Access, Fox, and Oracle databases, you must code your program with three different database languages. This can be a very difficult or time-consuming task.

This is where ODBC enters the picture. When programming to interact with ODBC, you only need to speak the ODBC language (a combination of ODBC API function calls and the SQL language). The ODBC Manager determines how to contend with the type of database you are targeting. Regardless of the database type you are using, all of your calls will be to the ODBC API. All you must do is install an ODBC driver specific to the type of database you are using.

Structured Query Language

Structured Query Language (SQL) is a standard language for defining and accessing relational databases. All of the major database servers used in client/server applications work with SQL. SQL is a query language designed to extract, organize, and update information in relational databases.

Each database vendor has its own particular dialect. These dialects are similar to one another, but different enough that programmers must pay close attention to which RDBMS is being used. The most important dialects of SQL are ANSI/ISO SQL, IBM DB2, SQL Server, Oracle, Ingres, and ODBC.

SQL uses statements to get work done. Examples of some of these statements are:

- SELECT
- INSERT
- DELETE
- UPDATE
- DECLARE
- OPEN
- CLOSE
- CREATE
- PREPARE
- DESCRIBE

SQL Call Level Interfaces

When a Call Level Interfaces (CLI) is used, a program requests database services by calling special SQL interface routines, rather than embedding SQL statements directly into the program. There are two distinct types of CLIs. First, each DBMS vendor provides its own unique API for its database. The vendor-specific API is usually the most efficient way to access the database, but each vendor's API is unique. As a result, if you decide to write programs that use a vendor API, you lock yourself into using that vendor's DBMS. However, your programs are efficient as possible.

The second type of CLI is a standard or open API, which is supported by more than one database vendor. Several open database APIs are available, one of which is ODBC. ODBC is a standard CLI for accessing SQL databases from Windows.

Genetic Analysis Technology Consortium

The Genetic Analysis Technology Consortium (GATC) was formed in an attempt to standardize the rapidly growing field of array-based genetic analysis. The consortium was created to provide a unified technology platform to design, process, read, and analyze DNA-arrays. The goal of the GATC is to make micro-arrays broadly available and provide a technology platform that allows investigators to use components from multiple vendors.

Databases and GeneSpring

Experiment data is not always stored on the researcher's desktop in simple text files. Sometimes the data are stored on a relational database. GeneSpring can save and load all types of data to an SQL database through ODBC.

Experiment data can be loaded from a database simply by telling GeneSpring which table(s) contain the data and which columns contain the experimental index. You load in the data using the Experiment Wizard almost exactly as you would if they were text files (see "Getting Data from a Database" on page A-22). The only difference is you enter experiment identifiers instead of file names, and SQL table columns instead of tab-delimited column headers.

Parameters describe what the database knows about each sample. Different databases have different ways of storing parameters, so they must be retrieved by explicit SQL statements. Silicon Genetics can provide these for GATC and help write these for individual databases. This needs to be done only once. Afterwards, the customer simply chooses the database and GeneSpring retrieves data from it. Normalization and other options can also be set for a database.

Adding an Experiment from a Database

This section describes the process of making your database visible to GeneSpring. The database administrator should have completed this task for you already. If this task has been completed, go to "Connecting your Database to GeneSpring" on page A-4.

Creating a New ODBC Source

To create a new ODBC source:

1. Select **Start** > **Control Panel** > **Administrative Tools** > **Open Data Sources (ODBC)**.

   The OCBC Data Source Administrator window opens.

2. Go to the system **DSN**.

3. Click **Add** for a new Create New Data Source window.

4. Select the correct type of database from the scrolling list. A new panel appears.

5. Give the experiment a name.

   Remember that experiment names in GeneSpring are case sensitive.

6. Click **Select** to browse for the correct database.

   In most cases, you will connect to a new computer (server) to access the database. The new entry appears in the list of databases.

Testing Your ODBC Connection

To test your ODBC connection:

1. Open **Excel**.
2. Select **Data** > **Get External Data**.
3. Select **New Database Query**. Look for your database in the presented list.

Connecting your Database to GeneSpring

The following section describes how to use the database configuration file. You must customize this file for your system before running GeneSpring. You can have any number of different configuration files

The purpose of this file is to tell GeneSpring how to read the database as if it were a simple text file. It pulls the data together and places it in columns recognized by GeneSpring. Column names and sample name references are entered in the Experiment Wizard as normal.

To use the database configuration file:

1. Using your file management software, create a new folder in the data directory of GeneSpring called Databases.
2. Create a file named dataloader.xml. Instructions for setting up this file are provided in the section below, *Configuration File Reference*.
3. Issue a SQL command to retrieve the parameters in all samples. Use MicroSoftQuery in Excel to generate SQL commands.
   a. In Excel open the **Tools** menu.
   b. Select **Get External Data**.
   c. Select **New Database Query**.
   d. Make sure you specify what to edit in MicroSoftQuery.

Configuration File Reference

The following section contains a list of the tags used in the database configuration file (`dataloader.xml`). The configuration file is in XML format and uses tags enclosed in angle brackets much like an HTML document.

In such a document, an *element* consists of a tag enclosed in angle brackets, and usually includes a closing tag. For example, the top-level element of this configuration file is the External Database Configuration element. This element consists of opening and closing tags, for example:

```
<ExternalDatabaseConfiguration>
...
</ExternalDatabaseConfiguration>.
```

An element's *contents* are tags or text nested between the current element's opening and closing tags. The following are examples of elements with contents:

```
<PhysicalDatabase>
    <UserName>BioMan</UserName>
</PhysicalDatabase>
```

In the above example, the `<UserName>` element (including its own contents and its closing tag) is the contents of the `<PhysicalDatabase>` element. The username `BioMan` is the contents of the `<UserName>` element.

Attributes are values defined within the opening tag of the element itself. In the following example, `name is an attribute, and "dbname"` is the value of the `<PhysicalDatabase>` element:

```
<PhysicalDatabase name="dbname">
```

An element may have any number of attributes or contents. An empty element (an element that has attributes but no contents) can be closed within the opening tag by adding a slash at the end, for example: `<tag value="example" />`.

Tag Reference

Table A-1 provides a list of all available tags for the database configuration file. For more detailed information on the use of each tag, see "Tag Definitions" on page A-8.

The Element column lists each of the available tags. The Contents column lists the type of contents that tag can contain (for example, plain text, or the names of the tags it can contain). The Attributes column lists the attributes that tag can contain.

Table A-1 Database Configuration File Tags

Element	Contents	Attributes
<ExternalDatabaseConfiguration>	<GeneralConfiguration> <Database>	n/a
<GeneralConfiguration>	<LoadClass> <ProcessedDataListFile>	n/a
<Database>	<PhysicalDatabase> <TechnologyType> <Header> <GenomeNames> <GetSampleIDs> <GetSampleAttributes> <GetFile> <GetRawData>	name, icon
<LoadClass>	plain text	n/a
<ProcessedDataListFile>	plain text	n/a
<PhysicalDatabase>	<UserName> <Password> <URL> <Prefetch>	name
<TechnologyType>	n/a	name

Table A-1 Database Configuration File Tags (Continued)

Element	Contents	Attributes
<Header>	<Author> <Research_Group> <Organization>	n/a
<GenomeNames>	<GenomeMappingSpec>	n/a
<UserName>	plain text	n/a
<Password>	plain text	n/a
<URL>	plain text	n/a
<Prefetch>	plain text	n/a
<Author>	plain text	n/a
<Research_Group>	plain text	n/a
<Organization>	plain text	n/a
<GetSampleIDs>	<DatabaseQuery> <DataDirectory> <FileNameMask> <IDFromFileName> <JavaQuery>	location
<GetSampleAttributes>	<DatabaseQuery> <JavaQuery>	cacheable numeric
<GetFile>	<DatabaseQuery> <JavaQuery>	type location deleteAfterwards mimeType
<GetRawData>	<DatabaseQuery> <Format>	n/a
<GenomeMappingSpec>	n/a	targetName sourceName baseDirectory
<DatabaseQuery>	SQL command	useGenomeName db
<DataDirectory>	plain text	n/a
<FileNameMask>	plain text	n/a
<IDFromFileName>	<RegexpMatch> <DatabaseQuery>	n/a
<RegexpMatch>	plain text	n/a>
<JavaQuery>	n/a	class extraArgs

Table A-1 Database Configuration File Tags (Continued)

Element	Contents	Attributes
<Format>	<GeneColumn> <Headlines> <SignalColumn> <NormalizedColumn> <ReferenceColumn> <SignalBackgroundColumn> <ReferenceBackgroundColumn> <ExperimentWorkedColumn> <ExperimentWorkedDesignation> <ExperimentAbsentDesignation> <ExperimentMarginalDesignation> <RegionColumn> <TreatNoSignalAsInvalid> <LowerBoundOnSignalColumn> <UpperBoundOnSignalColumn> <StandardDeviationSignalColumn> <ColumnHeaderLine>	type
<GeneColumn>	plain text	n/a
<Headlines>	plain text	n/a
<SignalColumn>	plain text	n/a
<NormalizedColumn>	plain text	n/a
<ReferenceColumn>	plain text	n/a
<SignalBackgroundColumn>	plain text	n/a
<ReferenceBackgroundColumn>	plain text	n/a
<ExperimentWorkedColumn>	plain text	n/a
<ExperimentWorkedDesignation>	plain text	n/a
<ExperimentAbsentDesignation>	plain text	n/a
<ExperimentMarginalDesignation>	plain text	n/a
<RegionColumn>	plain text	n/a
<TreatNoSignalAsInvalid>	plain text	n/a
<LowerBoundOnSignalColumn>	plain text	n/a
<UpperBoundOnSignalColumn>	plain text	n/a
<StandardDeviationSignalColumn>	plain text	n/a
<ColumnHeaderLine>	plain text	n/a

Tag Definitions

<ExternalDatabaseConfiguration>

The top-level element defining the entire database configuration. This element contains all of the other tags.

Contents: `<GeneralConfiguration>`, `<Database>`

Attributes: n/a

Usage:

```
<ExternalDatabaseConfiguration>
...
</ExternalDatabaseConfiguration>
```

Notes: Required, can appear only once in the configuration file

<GeneralConfiguration>

The element containing all of the general configuration options for the database.

Contents: `<LoadClass>`, `<ProcessedDataListFile>`

Attributes: n/a

Usage: `<GeneralConfiguration>...</GeneralConfiguration>`

Notes: Required, can appear only once in the configuration file

<Database>

The element containing the specifics of the source or sources of sample data, whether it is in a database or a directory of flat files. You must have one Database section for each source to which you will connect. The icon attribute is optional, and lets you specify a graphical image to represent the database.

Contents: `<PhysicalDatabase>`, `<TechnologyType>`, `<Header>`, `<GenomeNames>`, `<GetSampleIDs>`, `<GetSampleAttributes>`, `<GetFile>`, `<GetRawData>`

Attributes: name (required), icon

Usage:

<Database name= “Affymetrix Database” icon= “{/usr/local/graphics/icon.gif” >
...
`</Database>`

Notes: Required, can appear multiple times

<LoadClass>

Loads the driver that connects to the database. In some cases you may want to use a JDBC driver written in Java which must be instantiated at startup. You can specify any number of these drivers. Any class you specify, however, must be in your CLASSPATH. This element is optional. If you are using a default driver, this is not necessary, but if you are using a specific driver, you must specify it here.

Contents: plain text

Attributes: n/a

Usage: `<LoadClass>sun.jdbc.odbc.JdbcOdbcDriver</LoadClass>`

Notes: Optional, frequently used

<ProcessedDataListFile>

This setting specifies where the database will save the list of samples that have been uploaded.

Contents: plain text

Attributes: n/a

Usage: `<ProcessedDataListFile>/usr/database/ProcessedList.txt</ ProcessedListDataFile>`

Notes: Required

<PhysicalDatabase>

You must have one Physical Database tag for each physical SQL database to which you will connect. The “name” element specifies the database name for any `<DatabaseQuery>` tags that occur within the `<Database>` element. If you are loading data from flat files, you may not use the `<PhysicalDatabase>` element.

Contents: `<UserName>`, `<Password>`, `<URL>`, `<Prefetch>`

Attributes: name

Usage: <PhysicalDatabase name=”dbname”>...`</PhysicalDatabase>`

Notes: Required to retrieve files from a SQL database

<UserName>

This element specifies the username for logging into the sample database.

Contents: plain text

Attributes: n/a

Usage: `<UserName>`Ndege MacKenzie`</UserName>`

Notes: Required for `<PhysicalDatabase>`

<Password>

This element specifies the password for logging into the sample database.

Contents: plain text

Attributes: n/a

Usage: `<Password>dbPassword</Password>`

Notes: Required for `<PhysicalDatabase>`

<URL>

Specifies the physical address of the database or directory from which you will retrieve sample data.

Contents: plain text

Attributes: n/a

Usage: `<URL>jdbc:odbc:database</URL>`

Notes: Required for `<PhysicalDatabase>`

<Prefetch>

This is an optional, rarely-used element that lets you specify how many rows to retrieve from the database during the prefetch process. This may be useful in certain cases where there are performance issues. This element is contained by `<PhysicalDatabase>`.

Contents: plain text

Attributes: n/a

Usage: `<Prefetch>20</Prefetch>`

Notes: Optional, rarely used

<TechnologyType>

Sample special field technology to be set for each sample uploaded. This identifies the chip or technology used for the sample.

Contents: n/a

Attributes: name

Usage: <TechnologyType name=”Affymetrix”/>

Notes: Optional

<Header>

This element specifies header fields to be set for each sample uploaded from the current database.

Contents: `<Author>`, `<Research_Group>`, `<Organization>`

Attributes: n/a

Usage: `<Header>...</Header>`

Notes: Required for `<PhysicalDatabase>`

<Author>

Specifies the sample author to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: `<Author>`Than Nguyen`</Author>`

Notes: Required for `<Header>`

<ResearchGroup>

Specifies the research group to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: `<ResearchGroup>`Discovery Central</Research Group>

Notes: Required for `<Header>`

<Organization>

Specifies the organization to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: `<Organization>`Cures R Us`</Organization>`

Notes: Required for `<Header>`

<GenomeNames>

This setting lets you associate samples with genomes. One database may have several genomes. Within this element there must be at least one `<GenomeMappingSpec>` tag.

Contents: `<GenomeMappingSpec>`

Attributes: n/a

Usage: `<GenomeNames>...</GenomeNames>`

Notes: Required, can appear only once in a configuration file

<GenomeMappingSpec>

This element specifies the name of the genome in your sample source, the target genome on GeneSpring to upload it into, and the base directory to use for `<GetSampleIDs>` or `<GetFile>` elements. This tag must appear at least once. However, if the "useGenomeName" attribute is set to false in the `<DatabaseQuery>` tag for `<GetSampleIDs>`, this tag must appear *only* once. The attribute values for this tag are as follows:

- targetName—The genome on GeneSpring into which the samples will be uploaded
- sourceName—The name of the database or directory from which to retrieve samples
- baseDirectory—The directory to use for `<GetSampleIDs>` or `<GetFile>` elements if no data directory is specified

Contents: n/a

Attributes: targetName, sourceName, baseDirectory

Usage:

```
<GenomeMappingSpec targetName="Yeast" sourceName="YeastDB"
baseDirectory="/"/>
```

Please note in the above example that the `baseDirectory` attribute value is "/". The second slash (/) takes the place of the `</GenomeMappingSpec>` tag.

Notes: Required

<GetSampleIDs>

This element specifies the location from which to upload samples. There are three accepted values for the “location” attribute:

- database—perform a database search
- directory—locate files in a directory
- java—upload files based on the result of a Java call

These values are case-insensitive.

Contents: `<DatabaseQuery>`, `<DataDirectory>`, `<FileNameMask>`, `<IDFromFileName>`, `<JavaQuery>`

Attributes: location

Usage: <GetSampleIDs location= “database”>...`</GetSampleIDs>`

Notes: Required for `<PhysicalDatabase>`

<GetSampleAttributes>

Specifies parameters for retrieving attributes associated with samples. These can include the name, value, or units of the sample attribute, and possibly a flag specifying whether the attribute is numeric. The “cacheable” attribute defines whether to cache sample attribute values for previously-uploaded samples. This can greatly improve performance for uploads from external databases. This will not affect automatic upload performance, since in this case sample attributes are already retrieved only once. Acceptable values are “true” and “false”.

The “numeric” attribute indicates whether the retrieved values should be considered numeric. Acceptable values are “yes”, “no”, or “guess”. This attribute is used only for database queries. If the Java query option is used, this setting is overridden.

Contents: `<DatabaseQuery>`, `<JavaQuery>`

Attributes: cacheable, numeric

Usage: <GetSampleAttributes cacheable= “true” numeric= “guess”>...`</GetSampleAttributes>`

Notes: Optional

<GetFile>

Specifies parameters for retrieving associated files. This element has four attributes.

The “type” attribute specifies the type of file to be retrieved. This attribute is required, and is case-insensitive. There are five possible values:

- Sample Image—a picture or pictures of the biological sample
- Array Image—a picture or pictures of the scanned array(s)
- CEL File—an Affymetrix CEL file (actually stored as a general attachment with MIME type application/x-AffyCELFile)
- Raw Data File—a raw data file or files
- attachment—a general attachment

The “location” attribute specifies the location of the files to be retrieved. This attribute is required, and is case-insensitive. There are four possible values:

- database—returns the contents of the file (typically a blob)
- file—returns a file pathname
- URL—returns a URL
- java—returns com.signetics.ext.database.getFile

The “deleteAfterwards” attribute specifies whether to delete the file once it has been imported. Accepted values are “true” and “false”. This attribute applies only if the “location” attribute is set to “file”. This attribute is optional. If not specified, its value defaults to “false”.

The “mimeType” attribute specifies the MIME type of the file or files being retrieved. Any valid MIME type is an acceptable value. This attribute is optional.

Contents: `<DatabaseQuery>`, `<JavaQuery>`

Attributes: type, location, deleteAfterwards, mimeType

Usage: <GetFile type= “Sample Image” location= “database” deleteAfterwards= “true” mimeType= “image/gif”>...`</GetFile>`

Notes: Optional

<GetRawData>

This option specifies how to retrieve the actual sample data. This may come from either a database or a raw file (or files). The raw file itself may have been a file downloaded from a database or extracted from a Java class.

If the data is located in a database, use `<DatabaseQuery>` to retrieve it. If it is in a file or directory of files, it is interpreted as a tab-delimited file, and you must specify the file format using the `<Format>` tag.

Contents: `<DatabaseQuery>`, `<Format>`

Attributes: n/a

Usage: `<GetRawData>...</GetRawData>`

Notes: Required

<DatabaseQuery>

This element lets you enter a SQL query that produces a list of sample identifiers, attributes, or other data based on the provided genome name. If “useGenomeName” is true, the SQL query is passed the sourceName specified in the current `<GenomeMappingSpec>` tag. Use the “db” attribute to specify the database to query.

Accepted values for “useGenomeName” are “true” and “false”.

The data retrieved by this option varies depending on which tag contains it, as follows:

- `<GetSampleIDs>`—a list of sample identifiers
- `<GetSampleAttributes>`—three columns (or a multiple of three columns, in which case each set of three is considered independently). These three columns are:

- sample attribute value
- sample attribute name
- sample attribute units

Each row represents one attribute. If there is more than one set of three columns, for each set of three, each row represents an attribute.

- `<GetFile>`—if the “location” attribute is set to “database”, returns two or three columns:
  - the data
  - the filename
  - the mime type (if present, this overrides the mimeType specified in `<GetFile>`)

  Each row in the result represents a file to be loaded.

Contents: SQL command

Attributes: useGenomeName, db

Usage: <DatabaseQuery useGenomeName= “true” db= “dbname”>select ID from Experiments where Experiments.chipType=?`</DatabaseQuery>`

Notes: Required if the location attribute value in `<GetSampleIDs>` is “database”.

<DataDirectory>

If samples are contained in flat files rather than a database, this setting specifies the directory in which sample files are located. If a directory is not specified or does not begin with an absolute path, the baseDirectory attribute of the `<GenomeMappingSpec>` tag is used.

Contents: plain text

Attributes: n/a

Usage: `<DataDirectory>`/usr/share/affy`</DataDirectory>`

Notes: Optional

<FileNameMask>

FileNameMask is applied to all files in DataDirectory to filter the FileNames. If the DataDirectory is not specified or does not begin with an absolute path, the baseDirectory attribute of the current`<GenomeMappingSpec>` section is used. If baseDirectory does not begin with an absolute path, the current user directory is used.

Contents: plain text

Attributes: n/a

Usage: `<FileNameMask>`*/AffyChipID*.chip`</FileNameMask>`

Notes: Required for retrieving data from flat files in a directory

<IDFromFileName>

This lets you generate sample IDs directly from file names. If sample IDs are generated using <Regexpmatch>, only one genome should be specified in the <GenomeNames>/ <GenomeMappingSpec> tags. The result of the <RegexpMatch> on the file names provides the sample IDs. If you are using <DatabaseQuery> instead, the file names are passed as arguments to the specified SQL query.

Contents: plain text

Attributes: n/a

Usage: <IDFromFileName>...</IDFromFileName>

Notes: Optional

<RegexpMatch>

When using <IDFromFileName>, use either this tag *or* <DatabaseQuery>, but not both.

Contents: plain text

Attributes: n/a

Usage: <RegexpMatch>AffyChipID(.*)\.chip</RegexpMatch>

Notes: Optional

<JavaQuery>

lets you use a Java class to return an array of identifiers. You specify a command such as:

```
<JavaQuery class="com.pharma.database.getParameters"
extraArgs="Blah"/>
```

and it creates an instance of com.pharma.database.getParameters using the default constructor. That class should implement com.signetics.ext.database.GetAttributes. Then for each attribute, a function is called with arguments of the database identifier, the database genome name, and the extra argument. The return value is an array of com.signetics.ext.database.Attribute objects, each with name, value, units and isNumeric fields.

Contents: n/a

Attributes: class, extraArgs (optional)

Usage: <JavaQuery class="com.pharma.database.getIDs" extraArgs=""/>

Notes: Optional, applies only to <GetSampleIDs>, <GetSampleAttributes>, and <GetFile>.

<Format>

Specifies the format of raw data to be retrieved. If this data is in a known format, you can specify it using the “type” attribute. Currently supported types are:

- Incyte Internet Download
- Incyte
- Affymetrix
- Affymetrix Pivot Table
- AtlasImage
- GenePix Results
- Imagene
- ScanArray
- QuantArray

Important: Affymetrix Pivot files with more than one sample per file are not supported.

If you are retrieving data from a directory containing data in multiple formats, any files that do not match the format you specify here will be ignored.

If data are being retrieved from a database as a set of columns using a SQL query, a known format type may not be used and must be explicitly defined using the format described below.

If your data are not in a standard format, you must define the format using the available tags. Columns can be specified either as a number (first column=1) or header. If columns are specified by the header, and data is retrieved from a database using a SQL query, make sure the headers retrieved in the SQL query exactly match the headers specified here. It is a good idea to write your SQL queries as follows:

```
select column1 as “column1” from table_name where ...
```

If a column is not used, you can omit the line or enter -1 (“”for strings).

Contents: <GeneColumn>, <Headlines>, <SignalColumn>, <NormalizedColumn>, <ReferenceColumn>, <SignalBackgroundColumn>, <ReferenceBackgroundColumn>, <ExperimentWorkedColumn>, <ExperimentWorkedDesignation>, <ExperimentAbsentDesignation>, <ExperimentMarginalDesignation>, <RegionColumn>, <TreatNoSignalAsInvalid>, <LowerBoundOnSignalColumn>, <UpperBoundOnSignalColumn>, <StandardDeviationSignalColumn>, <ColumnHeaderLine>

Attributes: type

Usage: <Format type= “Affymetrix”/> **or** <Format>...</Format>

Notes: Required

<GeneColumn>

Specifies which column in the sample data contains the gene identifier. This tag is used only if your data are in a nonstandard format.

Contents: plain text

Attributes: n/a

Usage: `<GeneColumn>1</GeneColumn>`

Notes: Required if data type is not specified in the `<Format>` tag.

<Headlines>

Number of header lines to skip at the top before further processing. This can usually be determined automatically if the columns are specified by header. This tag does not apply when sample data are retrieved from a database.

Contents: plain text

Attributes: n/a

Usage: `<Headlines>`0</Headlines

Notes: Optional

<SignalColumn>

Specifies the column containing the raw signal data.

Contents: plain text

Attributes: n/a

Usage: `<SignalColumn>`31`</SignalColumn>`

Notes: Required if data type is not specified in the `<Format>` tag.

<NormalizedColumn>

Specifies the column containing normalized data.

Contents: plain text

Attributes: n/a

Usage: `<NormalizedColumn>`30`</NormalizedColumn>`

Notes: Optional, rarely used

<ReferenceColumn>

Specifies the column containing raw reference data. This is typically present in two-color experiments.

Contents: plain text

Attributes: n/a

Usage: `<ReferenceColumn>32</ReferenceColumn>`

Notes: Optional

<SignalBackgroundColumn>

Specifies the column containing the background signal to be subtracted from the main signal before further processing.

Contents: plain text

Attributes: n/a

Usage: `<SignalBackgroundColumn>-1</SignalBackgroundColumn>`

Notes: Optional, rarely used

<ReferenceBackgroundColumn>

Specifies the column containing the background signal to be subtracted from the reference signal before further processing.

Contents: plain text

Attributes: n/a

Usage: `<ReferenceBackgroundColumn>-1</ReferenceBackgroundColumn>`

Notes: Optional, rarely used

<ExperimentWorkedColumn>

Specifies the column containing a flag or flags indicating success of the measurement.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentWorkedColumn>33</ExperimentWorkedColumn>`

Notes: Optional, see also `<ExperimentWorkedDesignation>`, `<ExperimentAbsentDesignation>`, `<ExperimentMarginalDesignation>`

<ExperimentWorkedDesignation>

Specifies the flag in the column specified by `<ExperimentWorkedColumn>` that indicates that the measurement worked well.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentWorkedDesignation>P</ExperimentWorkedDesignation>`

Notes: Optional, see `<ExperimentWorkedColumn>`

<ExperimentAbsentDesignation>

Specifies the flag in the column specified by `<ExperimentWorkedColumn>` that indicates that the measurement did not work well.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentAbsentDesignation>A</ExperimentAbsentDesignation>`

Notes: Optional, see `<ExperimentWorkedColumn>`

<ExperimentMarginalDesignation>

Specifies the flag in the column specified by `<ExperimentWorkedColumn>` that indicates that the measurement worked only marginally.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentMarginalDesignation>M</ExperimentMarginalDesignation`

Notes: Optional, see `<ExperimentWorkedColumn>`

<RegionColumn>

Specifies the column that indicates regions to be normalized separately.

Contents: plain text

Attributes: n/a

Usage: `<RegionColumn>-1</RegionColumn>`

Notes: Optional, rarely used

<TreatNoSignalAsInvalid>

Specifies whether a signal of “0” should be treated as blank. If no value is specified for this tag, it defaults to “no”. Accepted values are “no” and “yes”.

Contents: plain text

Attributes: n/a

Usage: `<TreatNoSignalAsInvalid>no</TreatNoSignalAsInvalid>`

Notes: Optional, rarely used

<LowerBoundOnSignalColumn>

When an error model is known, specifies a lower bound on the signal value.

Contents: plain text

Attributes: n/a

Usage: `<LowerBoundOnSignalColumn>-1</LowerBoundOnSignalColumn>`

Notes: Optional, rarely used

<UpperBoundOnSignalColumn>

When an error model is known, an upper bound on the signal value.

Contents: plain text

Attributes: n/a

Usage: `<UpperBoundOnSignalColumn>-1</UpperBoundOnSignalColumn>`

Notes: Optional, rarely used

<StandardDeviationSignalColumn>

When an error model is known, specifies the standard deviation of the signal value.

Contents: plain text

Attributes: n/a

Usage: `<StandardDeviationSignalColumn>-1</StandardDeviationSignalColumn>`

Notes: Optional, rarely used

<ColumnHeaderLine>

Specifies the row containing the header names. Usually this can be determined automatically.

Contents: plain text

Attributes: n/a

Usage: `<ColumnHeaderLine>-1</ColumnHeaderLine>`

Notes: Optional, rarely used

Getting Data from a Database

To get data from a database:

- In the main GeneSpring window, select **File > Get Data from Database**.

The majority of the remaining Experiment Wizard panels are filled in automatically. If you left the debug setting ‘true’ an additional window opens. When the query boxes appear these will contain actual SQL commands.

GeneSpring must re-query the database each time you restart the program. If this takes too long, you can right-click the appropriate database icon and select the **Save to Disk** option. All commands in the *experiment* files can also be added to the *database* file.

Using Complex Databases

You can link various tables together in SQL. This procedure typically requires a proficient user of databases. Check with your database administrator if you have questions.

There are many ways to enter and organize data within databases. If the data organization in your database is complex, you might want to make separated tables for your data or part of your data. For example, you could make a separate table just for parameters, like the Table A-2.

Table A-2 Parameter Table

Sample 1	Parameter Name	Parameter Value
1	elephants	2
2	elephants	34
2	daises	30

In this table there are no parameters in the individual columns. All parameters tables should have an associated sample number.

If you use a GATC database, you must re-link all the sample numbers to the parameter numbers. In that case you must define an SQL. In that case, you must define a SQL line to get those parameters, for example:

```
SQLgetParameters : select
```

This statement should retrieve values and names of the parameter.

B

Name Restrictions for Data Objects

Table B-1 describes the naming restrictions for data objects.

Table B-1 Naming Restrictions for Data Objects

Data Object	Restrictions
Genome	• Cannot duplicate the name of a pre-existing local genome • Cannot be longer then 80 characters • Cannot be blank • Only the following characters are allowed in genome names, all others are disallowed: a-z, A-Z, 0-9, space, _().-& (no comma)

Gene Systematic Name	• Genes cannot duplicate the name of a pre-existing gene
	• Genes cannot be blank
	• The database column widths in Signet are:
	  • Systematic name: 256
	  • Common name: 200
	  • GenBank: 50
	  • Synonyms: 100
	• The leading character cannot be a pound sign (#)
	• Tabs and return characters are disallowed

Gene Annotation	• The names must be unique, they cannot duplicate the name of a pre-existing annotation • Tabs and return characters are disallowed • They cannot be blank • The leading character cannot be a pound sign (#) • The following characters are disallowed: “, ‘, <, >, & (commas OK) • 200 character limit • No leading or trailing white-space characters
Gene Annotation Value	• Tabs and return characters are disallowed • 4000 character limit • No leading or trailing white-space characters
Sample Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot duplicate another sample’s name (in the same genome)

Table B-1 Naming Restrictions for Data Objects (Continued)

Sample Attribute Name	• 1000 character limit • Name cannot be blank • No leading or trailing white-space characters • No returns or line breaks • Names must be unique within a sample
Attribute Unit	• 1000 character limit • No leading or trailing white-space characters • No returns or line breaks
Attribute Value	• No character limit • No leading or trailing white-space characters
Experiment Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another experiment’s name (in the same folder)
Experiment Interpretation Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another experiment interpretation’s name (in the same experiment)
Parameter Name	• Parameter names cannot be blank • 50 character limit • The names must be unique within an experiment
Parameter Unit	No character limit
Parameter Value	• Parameter values cannot be blank • 50 character limit • Numeric parameter values must be numbers • Logarithmic parameter values must be positive numbers
Gene List	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another gene list’s name (in the same folder)

B-2 Name Restrictions for Data Objects

Table B-1 Naming Restrictions for Data Objects (Continued)

Classification Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another classification’s name (in the same folder)
Condition Tree Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another condition tree’s name (in the same folder)
Gene Tree Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another gene tree’s name (in the same folder)
Array Layout Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another array layout’s name (in the same folder)
Pathway Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another pathway’s name (in the same folder)
Bookmark Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another bookmark’s name (in the same folder)

Table B-1 Naming Restrictions for Data Objects (Continued)

Expression Profile Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another expression profile’s name (in the same folder)
External Program Name	• 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof) • Cannot duplicate another external program’s name (in the same folder)
Script Name	• Name cannot be blank • Name cannot be in an internal conflict with the names we use internally for scripts: • Name cannot be “Advanced Filtering” • Name cannot start with “library” • Name cannot start with “/library” • Name cannot start with “\library” • 80 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof)
Folder Name	• 200 character limit • No slashes (forward or backward -- \ or /) • No returns or line breaks • No leading or trailing white-space characters • Cannot be “all genes” or “all genomic elements” (or any capitalization variation thereof since they are reserved words) • Cannot duplicate another folder of it’s object type’s name (if it exists in the same folder)
Signet User Name	• 30 character limit • Allowable characters in names are: • Letters (upper or lower case) • Digits • Space • Hyphen • Underscore Nothing else is allowed

Table B-1 Naming Restrictions for Data Objects (Continued)

Signet Group Name	• 30 character limit • Allowable characters in names are: • Letters (upper or lower case) • Digits • Space • Hyphen • Underscore Nothing else is allowed
Signet User Password	46 character limit
Signet Default User Directory Name	200 character limit
Signet E-mail address	50 character limit
Signet User's Full Name	50 character limit

C

Array Layout

The Array Layout view in GeneSpring lets you see a synthetic picture of the arrays used in the current experiment. This view is useful in identifying arrays that display local shifts in intensity due to problems in probe deposition, hybridization, washing, or blocking. To use the Array Layout view, you need to create an array layout file.

Creating an Array Layout

To create a layout file in GeneSpring, you need at least one file to tell GeneSpring general information about the array (size, shape, features, format, name, etc.). This file should end in the *layout* extension. You usually need another file describing exactly which gene goes where.

Array Layout Format

The format of the layout file is a series of lines. Order does not matter. Each line consists of a property, a colon, and a value.

For example, `property : value`. Blank lines and lines starting with a number sign (#) are ignored by GeneSpring. The following properties are allowed in the file. As always, GeneSpring is case-sensitive, so use the capitalizations as presented here:

- **Name—**Name of this layout, to appear in the Navigator window of GeneSpring.
- **Icon—**(optional) Path of a 16 x 16 gif file to appear next to the layout in the Navigator window.
- **VerticalSubArrays—**(optional, default 1) Number of rows of sub-arrays.
- **HorizontalSubArrays—**(optional, default 1) The number of columns of sub-arrays.
- **HorizontalPerSubArray—**Number of columns of dots in a sub-array.
- **VerticalPerSubArray—**Number of rows of dots in a sub-array.
- **VerticalDuplication—**(optional, rarely used) When dots are duplicated vertically, the number of copies.
- **HorizontalDuplication—**(optional, rarely used) When dots are duplicated horizontally, the number of copies.
- **CommonArrayType—**Format of the array.
  - **Q-X-Y**—The data file contains two columns. The first is a list of genes, the second is a set of three numbers separated by commas or hyphens. The first is the "sub-array" number, the second is the X-coordinate, and the third is the Y-coordinate. All

numbers start counting from 1. The sub-arrays are counted left to right, top to bottom. The second column can optionally be enclosed in quotation marks.

- **Q-R-C**—Same as "Q-X-Y", except the X and Y coordinates are swapped.
- **CLONTECH LNL**—There is no datafile. All genes have systematic names of the form "B4c" indicating where they are in the array. The first (capital) letter indicates which sub-array; the number indicated which column, and the lower case letter indicates which row.
- **CLONTECH LNNL**—Same as LNL, except there are two digits instead of one.

- **DataFileName—**Name of a datafile linking locations with gene names in format given by the CommonArrayType choice.

Once you have created the *layout* file, save it in the Array Layouts folder of the genome folder for which the layout pertains. For example, if you have not changed the defaults set-up of GeneSpring the path to the layout folder in the yeast genome is:

`c:\program files\ silicongenetics\genespring\data\demo chips\yeast\arraylayouts`.

Examples for Arrays

Here is an example for Pat Brown's yeast layout. The following is from a file *`Pat.layout`*:

```
Name : Pat Brown's Yeast Layout
# Icon: XXX.gif
VerticalSubArrays: 2
HorizontalSubArrays: 2
HorizontalPerSubArray: 40
VerticalPerSubArray: 40
VerticalDuplication: 1
HorizontalDuplication: 1
CommonArrayType: Q-X-Y
DataFileName: PatLocationList.txt
```

Following are the first few lines of the file `PatLocationList.txt`:

```
YHR007C "1,13,1"
YBR218C "2,13,1"
YAL051W "1,14,1"
YAL053W "2,14,1"
YAL054C "1,15,1"
YAL055W "2,15,1"
YAL056W "1,16,1"
```

Here is an example for a CLONTECH Array, from a file `Clontech.layout`:

```
Name: Clontech 588
# Icon: XXX.gif
VerticalSubArrays: 2
HorizontalSubArrays: 3
HorizontalPerSubArray: 14
```

```
VerticalPerSubArray: 14
VerticalDuplication: 1
HorizontalDuplication: 2
CommonArrayType: Clontech
```

Making an array can be a complicated process. Contact Silicon Genetics Technical Support at **1-866-SIG-SOFT** or **support@silicongenetics.com** for more information on this topic.

Glossary

Annotation Gene annotations denote any gene information beyond the expression level data.

Array A set of spots on a chip, typically expressed as a set of intensity measurements. An array generally has one sample. If all of the interesting genes fit onto one array, the terms array, chip and sample can be considered synonymous.

Array Layout A synthetic picture of genes on arrays. The Array Layout view can be used to check for gross slide related problems.

Chip The measurements from a glass slide containing DNA samples for microarray analysis.

Classification A grouping of genes by k-means or SOM clustering that is stored in the Classifications folder.

Cluster A collection of genes that have been grouped according to a certain criteria, such as similar mean expression values.

Colorbar The rectangle on the far right of the main GeneSpring window. The intensity of the Colorbar in GeneSpring indicates the reliability of the data for each gene.

Condition A grouping of one or more samples.

Condition Tree A dendrogram used to show the relationships between the expression levels of conditions.

Control An experiment data set that provides a comparison or contrast to experimental results.

Control Strength The quantity the raw value is divided by to get the normalized value. (see also expression strength)

Data Objects Any downloadable or uploadable items in GeneSpring, such as genomes, gene lists, classifications, etc.

Dendrogram A diagram showing hierarchical relationships, based on similarity between elements, for example, similarity of gene expression levels.

Experiment A group of conditions associated together under one name. This generally means they were all performed using a particular set of parameters.

Experimental Parameter A variable used to describe the condition or conditions during an experiment. A set of parameter values defines a single experimental parameter. When the word “parameter” is used alone, it usually refers to an experimental parameter.

Experiment Interpretation Tells GeneSpring how to treat and display your experiment parameters and how normalized values should be treated.

Experiment Specification Area The area under the Genome Browser that indicates which (if any) sub-experiment is being displayed, for example, a particular time point in a time series experiment.

Expression Production of mRNA through transcription of a DNA gene sequence.

Expression Level The amount of mRNA produced by a given gene under specific conditions.

Expression Profile Lines representing gene profiles that you draw in the Genome Browser. You can then search for genes matching that profile.

External Program Analysis programs outside GeneSpring which can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

Flag A flag is an additional measurement marker in your data set. It can be assigned as present, marginal, unknown, or absent.

Folders The yellow icons denoting the various directories where data are stored, for example, Gene Lists folder, Experiments folder, etc.

Gene List A list of genes based on some criteria.

Gene Tree Dendrograms used as a method of showing relationships between the expression levels of genes over a series of conditions.

Genome The set of all genes on a chip or array.

Genome Browser The area of a GeneSpring window containing a visual representation of genes.

Main Window The first GeneSpring window that appears after you open a genome, such as the default yeast genome window that appears after initially starting the program.

Measurement The smallest "unit" of data recognized by GeneSpring. These raw values can be seen in the upper right table in the Gene Inspector.

Navigator The left panel of GeneSpring windows containing data organized into folders.

Normalize The use of statistical methods to eliminate systematic variation in microarray experiments that can influence measured gene expression levels.

Pathways A graphical representation of the interaction between gene products in a biological system. Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context.

Parameter Value One of the possible values assigned to a variable.

p-value The probability that a certain statistic is equal or more extreme to the observed value when the null hypothesis is true. The null hypothesis is that the two samples are the same.

Regulatory Sequence The sequence upstream of a given gene to which regulatory enzymes bind, thereby determining the amount of expression of a particular gene.

Replicate A replicate can be multiple spots on the same array representing the same gene (also referred to as a copy), the same sample in more than one array, or a biological replicate that is equivalent samples taken from more than one organism.

Sample The measurements taken from one or more chips containing a single liquid sample, or the data generated from a biological object placed onto an array or set of arrays.

Slider A horizontal scrollbar at the bottom of the GeneSpring window that changes the display of genes from one sub-experiment to another, for example, in a time series experiment, the slider moves the displayed genes across the different time periods.

t-test T-tests calculate p-values which measure the significance of differential gene expression in each condition.

Trust A measure of reliability of the data.

Two-color Experiment An experiment where a control is used.

Variable A factor such as a disease, drug concentration, patient name, pipette number, time, the strain of organism tested, or who performed the experiment, etc. These variables allow you to look for meaningful patterns in you data and deal sensibly with replicate experiments.

Index

Symbols
"A" data flag 27
"M" data flag 27
"O" data flag 27
"P" data flag 27
"U" data flag 27

Numerics
3D scatter plot view
  change axes 26
  change color settings 28
  change features 27
  change legend options 29
  change lines to graph 27
  display centroid 29
  display options 25
  open 25
  use error bars 29

A
Absent flag 27
actions 10
  drag and drop 10
  mouse 10
Add Column window 43
Add Gene to Pathway window 13
Add Gene window 45
Add Link window 17, 23, 25, 31
Add New window 30
Add Standard Link window 27
Advanced Filtering window
  buttons 30
  described 29
  restrictions list 30
  restrictions table 30
Advanced Find Genes window 4
Affine background correction 19, 28
Affymetrix
  expression data 6
  file formats 1
  normalize data 24
  open genome or array 2
  probesets 12
Agilent 1, 7
Ambiguous Gene Identifiers window 15
ambiguous identifiers 15
Amersham 1, 7
animation controls 20
  hide 5
  show 5
annotation
  annotations retrieved 8
  Common Name 10
  copy gene list 9
  create custom 48
  Description 10
  duplicate identifiers 49
  edit 53
  filter 63
  filter gene list 9
  find gene 4
  GenBank Accession Number 10
  import from file 47
  import with Master Table of Genes Editor 46, 53
  long name problem 51
  Map 10
  map location problems 8
  Master Table of Genes 36
  retrieved from GeneSpider 41
  save gene list 10
  standard annotations 9
  Synonym 10
  Systematic Name 10
  update with GenBank 6
  update with LocusLink 6
  update with Silicon Genetics 5
  update with UniGene 6
Annotations menu
  Build Homology Tables 10
  Build Ontology 10
  Edit Genes and Annotations 10

GeneSpider 9
Update Annotations from GenBank 9
Update Annotations from LocusLink 9
Update Annotations from Silicon Genetics 9
Update Annotations from UniGene 9
Make Gene Lists from Annotations 10
Annotations Summary window 52
array layout
create 1
folder 13
format 1
array layout view
change color settings 41
change legend options 41
display options 41
open 41
attribute
add standard 63
color 19
copy 62
create 60
delete 62
delete standard 65
described 58, 16
edit standard 64
filter 50
import 20, 39, 59
inspect 22
paste 62
standard, described 63
axis
3D scatter plot view 26
bar graph view 13
change order of parameters 44
condition scatter plot view 54
graph by genes view 47, 48
graph view 10
interpretation 67
scatter plot view 21
Axon 1, 7

B

bar graph view
change axis 13
change color settings 15
change features 14
change legend options 15
change lines to graph 15
display options 13
open 13
Benjamin and Hochberg False Discovery Rate 13
BioDiscovery 1, 8
blocks view
change color settings 8
change features 8
change legend options 8
display options 7
open 7
Bonferroni 12
multiple test correction 13
Step Down 13
Step Down (Holm) multiple testing correction 13
bookmark
create 3
folder 13
load 4
Build Homology Tables window 34, 23
buttons 25
Select Genomes Navigator 24
building block
Boolean 24
Clustering 31
Correlations 32
Count Groups 31
described 20
external program 15, 47
External Programs 37
Filter Gene Tools 34
Gene List Manipulations (Venn Diagram) 25
input 16, 21
Look Up 28
Make Groups 29
Merge-Split Groups 28
Numbers 30
open 38
open pre-defined 38
output 16, 21
Regulatory Sequence Search 35

save 43
Select Groups 29
Select with Boolean 24
Signet Downloading (Default Directory) 26
Signet Downloading (Specified Directory) 26
Signet Publishing (Default Directory) 27
Signet Publishing (Specified Directory) 27
Standardize Inputs 37
Statistical Analysis (ANOVA)
Statistical Analysis (ANOVA) building block 36
use from external programs 39

C

Change Colors window 68
change correlation 17, 14, 20
Change Owner window 10
Change Projects window 24, 29, 3
Choose Classification Name window 18
Choose Experiment Name window 22
Choose Output Directory window 14
Choose Type of Input window 51
Choose Type of Output window 52
Choose Zip to Import window 5
class prediction
described 29
interpret results 32
k-nearest neighbors 29
perform k-nearest neighbors 30
perform support vector machines 34
support vector machines 30
technical details 33
Class Prediction window 30
classification
color 19, 66
create 20
inspect 38
make gene list 19, 21
Classification Inspector window
buttons 39
Details table 39
open 37
Summary information section 38
table 39
Clontech 2, 9
clustering
condition tree 4
described 1
gene tree 3
k-means 2
perform condition tree 11
perform gene tree 8
perform k-means 5
perform QT 17
perform self-organizing map 14
QT 5
self-organizing maps 4
Clustering window 2
color
background 18
customize 19
downregulated 18
no color 18
normal 18
presets 18
structure 18
text 18
upregulated 18
Colorbar
change range 63
color by classification 66
color by parameter 65
color by significance 64
color by Venn diagram 64
described 19
expression 61
no color 65
save image 3
set expression 62
set trust 62
trust 61
Colorbar menu
Color by Classification 7
Color by Expression 7
Color by Parameter 7
Color by Significance 7
Color by Venn Diagram 7
Clear This Selection 7
Make Gene Lists 7
No Color 7
Column Editor window

described 6
open 9
set columns 10
set titles 9
compare genes to view
display second list 45
open 45
Complex Correlations window
Correlations table 13
Graph panel 12
Preview pane 12
condition
described 30
exclude in filtering 6
inspect 30
Condition Inspector window 30
buttons 31
Condition Correlations tab 32
described 30
Graph tab 33
open 30
Pictures tab 33
Sample and Parameters tab 31
Summary Information section 31
condition scatter plot view
change axis 54
change color settings 56
change features 55
change legend options 56
change lines to graph 55
display options 53
open 53
use error bars 56
condition tree clustering
perform 11
save results 12
view results 13
Conditions to Filter window 9
Configure Columns window 43
Confirm Gene List Folder Name window 32
conjectured regulatory sequence
described 7
make gene list 9
view 9
Conjectured Regulatory Sequence window 7
Copy Annotated Gene List window 6
correlation
change 14
described 11
distance 14
experiment weight 19
find similar gene 23
Pearson 14
perform 21
smooth 14
Spearman 14
Spearman confidence 15
Spearman two-sided 15
standard 14
upregulated 14
Create New Experiment window
assign to project 50
described 25
filter methods 26
filter results buttons 27
sample lists 26
selected samples buttons 27
Cross-Gene Error Model
described 73
enable 76
technical details 77
use 76
uses 75
Cross-Gene Error Model window 73
Custom Annotation window 48

D

data preprocessing 13
data restrictions
control signal 4
normalized data 4
raw data 4
data transformation
described 10
dye swap 11
log to linear values 11
real-time PCR 10
SAGE 10
set measurements 11
subtract background 10
data types
interpreted 7

normalized 7
raw 7
database
add experiment from 3
Call Level Interfaces 2
configuration file 4
connect to GeneSpring 4
described 1
GenBank 2
Genetic Analysis Technology Consortium 2
get data from 22
Import Data command 2
Import Data from Database wizard 23
LocusLink 4
NCBI 1
Open Database Connectivity 1
Preferences command 4
Silicon Genetics 1
Structured Query Language 2
tag definitions 8–21
tag reference 5–7
Test Connectivity command 11
UniGene 4
Update Annotations from GeneSpider command 9
use complex 22
d-Chip 7
defaults 25
browser settings 20
color settings 17
correlation settings 24
data files 16
database settings 17
firewall settings 20
group color settings 19
language settings 25
local settings 23
remote settings 23
search settings 24
Signet settings 22
standard color settings 18
system settings 21
Define Other Genomic Elements window 59
buttons 62
described 59
Filter on Annotation tab 61
Filter on Gene List tab 60
filtering options 60
Type a List tab 60
distance correlation 14, 18
drag-and-drop actions 10
duplicate identifiers
annotation 49
dye swap transformation 11

E

EB submissions 44
Edit Annotation Value window 54
Edit Attribute Value window 21
Edit Attributes window 59, 60, 62
Edit Link window 21, 25, 29, 34
Edit menu
Advanced Find Gene 4
Copy 4
Copy Experiment 4
Copy Gene List 4
Edit Gene List 4
Filter Navigator 4
Find Gene 4
Find Next 4
Inspect Selected Gene 4
Paste 4
Paste Classification 4
Paste Experiment 4
Paste Gene List 4
Preferences 4
Search 4
Standard Attributes 4
Undo 4
Eisen-like tree view 31
Electronic Signature window 5
Entrez mirror 25
error
ambiguous identifiers 15
external program 47
long name problem 11, 51
non-unique identifiers 6, 11, 14
normalization warnings 9
script 46
error bars
3D scatter plot view 29

condition scatter plot view 56
graph view 12
scatter plot view 24
Euclidean distance 18
Exclude Conditions window 6, 24
experiment
ambiguous identifiers 15
assign project 50
change interpretation 72
copy 30, 35, 42
create 28
create attribute 60
create parameter 41
delete interpretation 73
delete parameter 42
extract subvalue 43
filter on attribute 50
filter on keyword 51
filter on parameter
parameter
filter 48
filter on project 53
find interpretation 73
find similar sample 53
fold change 68
folder 12
import attribute 20, 39, 59
import parameter 39
import selected files 14
inspect 25
interpretation 27
interpretation, default 67
interpretation, described 65
load from file 3
log of ratio 68
normalization 28
open interpretation 71
parameter 37, 38, 27
paste 30, 34, 42
preprocess data files 13
save 23
select control files 16
select samples with multiple files 17
select signal files 16
set up columns 9
weighted correlation 19
Experiment Inspector window 25
Associated Files tab 29
Colorbar tab 28
Interpretations tab 27
Normalizations tab 28
open 25
Summary information section 26
Experiment Interpretation window
Cross-Gene Error Model 73
described 66
fold change 68
log of ratio 68
open 71
vertical axis modes 67
experiment link
add 23, 31
described 22
edit 25, 34
remove 25, 34
set order 24, 33
Experiment Link Order window 24, 33
Experiment Normalizations window 2
Experiment Parameters window 39, 42, 43, 44
Experiments menu
Create New Experiment 7
Cross-Gene Error Model 7
Duplicate Experiment 7
Experiment Interpretation 7
Experiment Normalizations 7
Experiment Parameters 7
Sample Manager 7
Experiments to Cluster window 19
export
gene list 5
image 1
MAGE-ML 10, 13
zip file 6
Export Genome window 6
Export Options window 2
expression
color genes by 62
described 61
profile 21
external program
add arguments 54
add custom delimiters 54

add input 51
add output 52
building block 47
create 50
described 47
error 47
folder 13
interface 47
remove argument 55
External Program Inspector window
open 59

F

Failed flag 27
False Discovery Rate 4
file
embl 4, 14
FASTA 13
gbk 7, 4, 14, 38, 9
Imagene 16
layout 1
MAGE-ML 1
pict 2
png 2
seq 12, 13
xml 1
zip 36, 1, 2, 3, 4, 5, 6, 8, 12
file formats 1
File menu
Bulk Upload to Signet 3
Close 3
Genome Manager 3
Import Data 2
Import Data from Database 2
Import GeneSpring Zip 3
Import Genome 3
Load Bookmark File 3
Log in to Signet 3
New Linked Window 3
New Pathway 3
New Program 3
New Window 3
Open Genome or Array 2
Affymetrix 2
Demo Chip 2
Print Image 3
Quit 4
Save Bookmark 3
Save Image 3
View Projects 2
View Not Assigned 2
filter
create advanced filters 31
exclude condition 6
on annotation 63
on arbitrary files 25
on attribute 50
on attributes 16
on confidence 12, 14
on data files 23
on error 10
on experiment 48
on expression level 5
on flags 21
on fold change 7
on gene list 62
on gene list numbers 27
on keyword 51
on parameter 48
on parameters 16, 18
on projects 33
project 53
save advanced filters 32
save as script 32
Filter Navigator window 41, 34
Advanced Filter Navigator tab 43
described 41
Filter Navigator by Keyword tab 42
filter projects 34
open 45
Filter on an Arbitrary File window 25
Filter on Confidence window 14
Filter on Data File window 23
Filter on Errors window 10
Filter on Expression Level window 5
Filter on Flags window 21
Filter on Fold Change window 7
Filter on Gene List Numbers window 28
Filter on Parameter window 18
Filtering menu
Advanced Filtering 8
Filter on Arbitrary File 8

Filter on Confidence 8
Filter on Data File 8
Filter on Error 8
Filter on Expression Level 8
Filter on Flags 8
Filter on Fold Change 8
Filter on Gene List Numbers 8
Filter on Parameters 8
find
annotation 4
by keyword 7
data by project 9
gene 1, 12
gene in view 3
pan in view 3
use advanced search criteria 11
zoom in view 2
Find Gene in View window 54, 3
Find Potential Regulatory Sequence window 2
Find Significant Parameters Results window
Genes tab 44
Parameters tab 43
Find Significant Parameters window 37, 39, 41
find similar
gene 14
sample 53
sample results
experiment
find similar sample results 57
Find Similar Genes window 23
Find Similar Samples Results window 57
Find Similar Samples window 54
Find Target Gene window 55
firewall, preferences 20
Fisher's Exact Test 33
flag
described 8, 27
precedence 28
types of 27
fold change 68
folder
Array Layouts 13
Bookmarks 13
Classifications 13
Condition Trees 13
Experiments 12
Expression Profiles 13
External Programs 13
Gene Lists 12
Gene Trees 13
Pathways 13
Scripts 14
font settings 25
Friedman 22

G

GenBank
database 2
update annotations 6
gene
adding in Master Table of Genes Editor 45
change display 8
deleting in Master Table of Genes Editor 46
find 9, 12
find annotation 4
find in view 54, 3
find similar 9
find with Genome Browser 1
import with Master Table of Genes Editor 46
inspect 9, 16
select 9
select multiple in Genome Browser 14
select single in Genome Browser 14
weight on intensity 56
Gene Inspector
open 14
Gene Inspector window 16
open 16
gene link
add custom 17, 25
add standard 19, 27
described 17, 25
edit 21, 29
remove 21, 29
set order 20, 28
Gene Link Order window 20, 28
gene list
color 19
copy 6
copy annotated 9
delete 8

described 1
display 2
edit 5
export 5
filter 62
filter on 8
filter on annotations 9
folder 12
generated p-values 27
inspect 34, 14
make 9, 14
make from annotations 30, 16
make from classification 19
make from conjectured regulatory sequence 9
make from expression profile 21
make from pathway 15
make from selected genes 21
make from Venn diagram 18
open 5
paste 6
save 6
save annotated 10
search preferences 15
show all genes 10
Gene List Editor window
buttons 3
Filter on Annotation tab 5
Filter on Gene List tab 4
Filter options 3
Filter Results table 3
Type a List tab 4
Gene List Inspector window 34
Associated Files tab 37, 14
buttons 35, 12
Gene List tab 35
Gene Lists tab 12
Genome Browser 35
Graph panel 12
open 33
Similar Lists tab 36, 13
Summary Information section 34, 11
Gene Ontology 9
gene tree clustering
perform 8
save results 10
view results 13
GeneSpider
errors 8
limitations 5
retrieved annotations 41
Update Genome window 7
GeneSpring
install 2
license key 3
memory usage 3
requirements 1
start 2
upgrade 4
workflow 5
GeneSpring window
Annotations menu 9
Array Layouts folder 13
Bookmarks folder 13
Classifications folder 13
Colorbar 19
Colorbar menu 7
Condition Trees folder 13
described 1
Edit menu 4
Experiments folder 12
Experiments menu 7
External Programs folder 13
File menu 2
Filtering menu 8
Gene List folder 12
Gene Lists folder 12
Gene Trees folder 13
Genome Browser 17
Help menu 10
Image window 20
Navigator 11
Navigator commands 14, 15
Pathways folder 13
Scripts folder 14
Tools menu 8
Window menu 10
GenMAPP 11
genome
add other elements 65
build homology table 34
build ontology 32

create custom 4, 7
create experiment links 23
create web links 15
delete 9
described 6, 1
edit properties 21
import from EMBL 4
import from GenBank 4
import from Silicon Genetics 2
import genomic sequence 22
import KEGG pathway 32
import sequence 12
import standard 2
import tab-delimited file 7
long name problem 11
name 27
open 13, 4
open inspector 20
other elements 59
pre-made 2
rename 8
select annotation 7
set default 16
update annotation 29
view history 21
Genome Browser
button
Show All Genes 18
Zoom Fully Out 18
Zoom In 18
Zoom Out 18
buttons 18
commands 17
described 16
find gene 1
menu
Delete Entire Tree 18
Display Entire Tree 18
Display Options 17
Display Parent of Subtree 18
Display Subtree 17
Find Genes Which Could Fit Here 18
Hide Associated Values 18
Load Sequence 17
Make List from Selected Genes 17
Make List from Subtree 17
Make Subtree 17
Remove Secondary Gene List 18
Reset Orientation 18
Save Bookmark 17
Save Expression Profile 17
Save Image 17
Select Genes Within Lines 18
Show Associated Values 18
Unsplit Window 18
Zoom Fully Out 17
Zoom In 17
Zoom Out 17
pan 3
select multiple genes 14
select single gene 14
zoom 2
Genome Inspector
described 1
edit gene properties 21
import genomic sequence 22
open 20
view history 21
Genome Inspector window
buttons 12
described 10
GenBank/EMBL Files tab 16
Genome Properties tab 12
History tab 13
Homology Tables tab 20
Sequence tab 17
Summary information 11
Web Links tab 19
Genome Manager
add folder 6
commands 3
delete folder 7
delete genome 9
described 1
export zip file 6
import zip file 5
open 4
rename folder 7
rename genome 8
Genome Manager window
buttons 2
commands 3

described 2
Navigator 2
GO SLIMS 31, 9
Golub method 33
graph by genes view
change axis 47, 48
change color settings 49
change features 49
change legend options 49
open 46
graph view
change axis 10
change color settings 12
change features 11
change legend options 12
change lines to graph 11
display options 9
open 9
use error bars 12

H
Harvard 2
Help menu
About 11
Email Technical Support 11
Keyboard Shortcuts 10
Show Helpful Hints 11
Signet Database 11
Silicon Genetics Online 11
Support and Training Resources 11
System Monitor 11
Test Database Connectivity 11
Tutorial 10
Update Full GeneSpring 10
User Manual 10
Version Notes 10
Holm 13
homology
build 34
homology table
build 23, 26
delete 28
described 23
supported organisms 25
view 27

I
icon legend 19
image
export 1
print active window 4
print Colorbar 4
print entire window 5
print Genome Browser 4
save active window 3
save Colorbar 3
save entire window 4
save Genome Browser 1
save Venn diagram 3
Image window 20
Imagene 16
import
annotation from file 47
build homology table 34
build ontology 32
create experiment links 23
create web links 15
custom genome 4, 7
experiment 3
experiment data 9
genome annotation 7
genome from EMBL 4
genome from GenBank 4
genome from Silicon Genetics 2
genome sequence 12
genomic sequence 22
KEGG pathway 32
long name problem 51
name genome 27
options 2
selected files 14
standard genome 2
tab-delimited genome file 7
update genome annotation 29
wizard 2
zip file 5
Import Annotations from File window 47
Import Attributes window 60
Import Data
Column Editor window 4
Create Experiment window 21
Define File Format and Genome window 3

Merge Files window 17
Preprocess Files window 5
Selected Files window 5, 13, 14
Import Genome
Select GenBank Files window 5
Import Genome window 2, 4
Import KEGG Pathways window 33, 11
Import/Update Sequence window 22
Incyte 2, 8
input
arrange 43
convert to knob 43
described 4
inspect
classification 38
condition 30
experiment 25
gene 16
gene list 34
interpretation 27
normalization 28
parameter 27
sample 20
Inspector
Classification 21
Condition 21
Experiment 21
External Program 21
Gene 21
Gene List 21
Genome 22
Inspect Selected Gene 22
Interpretation 22
Sample 22
Script 22
inspector
Genome 20
install
GeneSpring from CD 2
GeneSpring from web 2
intensity-dependent normalization 13
interpretation
change 72
default 67
delete 73
described 7, 65
find 73
fold change 68
inspect 27
log of ratio 68
open 71

K
KEGG 11
keyboard 10
keyboard shortcuts
graphics 12
menus 11
tools 11
tree view 12, 33
views 11
k-means clustering
classifications 3
described 2
perform 5
save results 7
view results 8
k-nearest neighbors
class prediction 29
perform 30
knob
convert to input 41
define 41
described 5, 16, 21
Kruskal-Wallis 15

L
language settings 25
license key 3
Load Bookmark window 4
LocusLink
database 4
update annotations 6
log of ratio 68
log to linear transformation 11
LOWESS normalization 13

M
Macintosh
keyboard tips 13
MAGE-ML
CompositeSequence identifier 11

export 10, 13
prepare for publication 11
MAGE-ML Export window 13
magnification 33
hide 5
show 5
zoom in 9
zoom out 9
Make Gene List from Annotations window 31
Make Gene Lists from Annotation window 16
Marginal flag 27
Master Table of Genes 36
Master Table of Genes Editor
add genes 45
annotations 41
buttons 39
cells 39
columns 37
commands 41
configure columns 43
copy text 58
custom annotation 48
delete genes 46
described 1, 36
EB submissions 44
edit annotation 53
edit columns 42
find gene 54
find text 57
import annotation 53
import annotations 46
import genes 46
open 42
save 58
sort data 45
summary box 37
table 37
median polishing 23
menu
Annotations 9
Colorbar 7
Edit 4
Experiments 7
File 2
Filtering 8, 2
Genome Browser 17
Help 10
Navigator 14
Navigator commands 15
Show 14, 33
Tools 8
View
GeneSpring window
View menu 5
Window 10
multiple testing correction
Benjamin and Hochberg False Discovery Rate 13
Bonferroni 13
Bonferroni Step Down (Holm) 13
described 12
multiple testing corrections
Benjamin and Hochberg False Discovery Rate 13
Bonferroni 12
Bonferroni Step Down 13
described 12
Holm 13
Westfall and Young 13

N

Name New Gene Tree window 10
Name Problems window 11, 51
name restrictions 1
Navigator
Array Layouts folder 13
Bookmarks folder 13
Classifications folder 13
commands 14, 15
Condition Trees folder 13
described 11
Experiments folder 12
External Programs folder 13
Gene List folder 12
Gene Lists folder 12
Gene Trees folder 13
hide 5
menu
Add New Folder 14
Assign Project 14
Attachments 14
Change Owner 14

Clear 14
Copy from Signet 14
Delete 14
Display 14
Display as Second Gene List 15
Draw Gene Profile 16
Duplicate 15
Edit 15
Export as MAGE-ML 15
Export as Zip 14
Help 15
Import Zip 14
Inspect 14
Make Gene Lists 15
Re-Annotate 15
Rename 14
Run 15
Save DB Data Locally 15
Set as Default 15
Set Secondary Experiment 15
Split Window 15
Translate 15
Upload to Signet 14
Use as Coloring 15
Use to Color Branches 15
Venn Diagram 15
Pathways folder 13
Scripts folder 14
show 5
negative control strengths 28
New 49
New Attribute window 20, 61, 63
New Correlation window 22
New Experiment Checklist window 24
New External Program window 48
New Gene List window 18
New Genome Checklist window 28
New Parameter window 41
New Scenario window 8
non-linear normalization 13
non-parametric test 17
Kruskal-Wallis 15
Student-Newman-Keuls 15
two-way ANOVA 22
non-parametric test statistic 15
non-unique identifiers error 6, 11, 14
Normalization Errors window 22
Normalization window
described 35
one-color experiments 36
two-color experiments 36
normalize
Affine background correction 19
Affymetrix data 24
apply defaults 6
assumptions 3
described 7, 1
editing steps 5
intensity-dependent 13
LOWESS 13
methods 3
non-linear 13
normalize data 4
per chip 15
per gene 20
per spot 12
pre-normalized values 12
region 19, 25
removing steps 6
reordering steps 6
repeated measurements 26
specific samples 4
two-color microarray data 25
viewing step descriptions 7
warnings 9

O

One-color experiments 36
one-way ANOVA
described 2
interpret results 23
perform 5
technical details 9
ontology
build 32, 9
described 32
ordered list view
change color settings 40
change features 39
change legend options 40
display options 39
open 39

Other Genomic Elements window
filter on annotation 63
filter on gene list 62
open 62
output
arrange 43
name 43

P
Packard Biochip ScanArray 8
Pairwise 2
pan 3
parameter
change order on axis 44
color 19, 65
color code 71
continuous 70
create 41
delete 42
described 7, 16
experiment 37
experiment interpretation 38
extract subvalue 43
find significant 37
find significant technical details 45
find significant with ANOVA 37
find significant with association test 39
find significant with correlation to parameters test 41
hidden 71
import 39
inspect 21, 27
interpret find significant results 43
non-continuous 70
use 37
value 37
Parameter Value Order window 44
parametric test 7
assume variances equal 17
don’t assume variances equal 17
Student-Newman-Keuls 15
Tukey 14
two-way ANOVA 18
use all available error estimates 17
pathway
add genes 12
commands 15
described 10
folder 13
GenMAPP 11
import image 12
import KEGG 32, 11
KEGG 11
make gene list 15
pathway view
add genes 43
change color settings 44
change features 44
change legend options 44
delete genes 43
display options 43
open 42
Pearson correlation 18, 14, 16
per-chip normalization 15
normalize to constant value 19
normalize to median 16
normalize to percentile 16
normalize to positive control genes 17
normalize to region 19
per-gene normalization
divide by specific samples 20
median polishing 23
per-spot normalization 12
physical position view
change color settings 19
change features 18
change legend options 19
display options 17
load regulatory sequence 19
open 17
plugin
create 68
data preprocessor 13, 63
described 63
edit 71
interactive 64
save 71
post hoc test
described 13
interpret results 23
Kruskal-Wallis 15
Student-Newman-Keuls 15

Tukey 14
Preferences window
Browser tab 20
Colors tab 17, 67
Computation tab 23
Data Files tab 16
Database tab 17
Firewall tab 20
Miscellaneous tab 24
open 15
Signet tab 22
System tab 21
preferences, setting 15
pre-normalized data 36
preprocessing, data files 13
principal component analysis
described 21
interpret results 26
on conditions 23
on genes 21
view results in ordered list 27
view results on conditions 24
view results on genes 22
view results on scatter plot 26
Principal Components Analysis window 21
print
active window 4
Colorbar 4
entire window 5
Genome Browser 4
probesets, Affymetrix 12
project
assign data objects 46
assign to experiment 50
described 40
expression profile 21
filter 53, 34
Filter Navigator window 34
find data 9
use folders 48
use inspector 48
use sample manager 49
projects
filtering on 33
p-test p-value 17

Q
QT clustering
described 5
perform 17
save results 18

R
real-time PCR transformation 10
region normalization 19, 25
regulatory sequence
described 1
enter 4
find 2
interpret results 6
view results 6
regulatory-compliant server, 5
Rename Column window 44
repeated measurements 26
Replace window 42, 57
replicate
described 7, 16
requirements
Macintosh 1
memory 3, 2
UNIX 2
Windows 1

S
SAGE transformation 10
sample
attribute 58
color 19
copy 62
create attribute 60
delete 62
described 45
filter on attribute 50
filter on experiment
filter
on experiment 48
filter on keyword 51
filter on parameter 48
filter on project 53
find similar 53
find similar results 57
import attribute 20, 59

inspect 20
list 46
paste 62
select multiple files 17
Sample Inspector window 20
open 19
Summary information section 20
Sample Manager window
assign to project 49
described 45
Filter on Experiment tab 48
Filter on Projects tab 53
filter results buttons 46
filtering methods 46
Filtering on Attributes tab 50
Filtering on Keyword tab 51
Filtering on Parameter tab 48
open 47
sample list 46
selected samples buttons 47
Save Bookmark window 3
Save Expression Profile window 22
Save Gene List window 6
Save New Classification window 7, 15
Save New Condition Tree window 12
Save New Experiment window 23, 50
Save New Genome window 4, 27
Save to File window 58
scatter plot view
change axis 21
change color settings 23
change features 22
change legend options 24
change lines to graph 23
display options 21
open 21
use error bars 24
script
2-fold Expression Change 2
2-fold Expression Change AND Filter on Noise NOT Input 2
Best k-means 2
Clustering 2-fold Change List 2
convert input to knob 43
convert knob to input 41
create 40
data file restrictions 7
define knob 41
described 1, 40
edit 20
Filter on Noise 2
folder 14
icon legend 19
input 4, 16, 21
inspect 14
invalid character error 12
knob 5, 21
long name error 12
move 46
open 6
output 16, 21
Pairwise Comparison 2
predefined 2
preferences 3
Probe Entire Enterprise Repository for Similar Conditions (PEER-C) 2
run locally 6
run locally, described 4
run remotely 11
run remotely, described 9
save 43, 45
Select k-means 2
Send Clustering Results to Signet 2
Series of k-means (increments of 5) 3
upload to Signet 13
view remote job 11
warning message 46
Script Editor window
Block panel 18
Browser panel 17
Icon legend 19
Navigator panel 17
Notes panel 18
Script section 18
search
Boolean 55, 63, 5, 7, 13, 42, 30, 31, 13, 9, 24
wildcard 51, 52, 56, 64, 5, 8, 11, 13, 43, 10, 25
Search window 12
Advanced Find Data tab 11
Find Data by Keyword tab 7

Find Data by Project tab 9
Find Gene tab 12
select
multiple genes in Genome Browser 14
single gene in Genome Browser 14
Select a Normalization Scenario window 9
Select Corresponding Files window 16
Select Data to Upload window 8
Select Groups for 1-Way Tests window 6
Select Groups for 2-Way Tests window 16
Select One Tab-delimited File window 47
Select Sample Pool window 56
Select Target Sample window 55
self-organizing map
clustering 14
described 4
save results 15
view results 16
set measurements transformation 11
Show menu 33
All Data 14
Filter 14
Not Assigned 14
Signet
add server 3
bulk data upload 7
change data ownership 10
delete server 4
described 1
edit server 4
log in 4
preferences 2
upload data 5
Signet Login window 4
Signet Upload window 5
Silicon Genetics
database 1
update annotations 5
similarity measures
change correlation 17, 14
clustering 20
distance 14
Pearson correlation 18, 14
smooth correlation 17, 14
Spearman confidence 15
Spearman correlation 18, 14
standard correlation 17, 14
technical details 15
two-sided confidence 15
upregulated correlation 18, 14
smooth correlation 17, 14, 19
socket
building block 20
described 16
input 43
output 43
Spearman
confidence 18
correlation 18, 14, 15, 17
two-sided confidence 18
spreadsheet view
copy 51
copy rows 51
find gene 51
open 50
select data 50
sort rows 51
Standard Attributes window 64
standard correlation 17, 14, 16
standard list
create 35
described 24, 35
use in searches 24, 35
Standard Web Links window 19
Statistical Analysis (ANOVA)
described 1
Statistical Analysis (ANOVA) window
choose experiment button 4
Choose Gene List button 4
P-value cutoff option 4
multiple testing corrections 4
Multiple testing corrections options 4
Parameter to Test option 4
post hoc tests 4
Student’s t-test 17
Student-Newman-Keuls 15
subtract background transformation 10
support vector machines
described 30
perform 34

T
Tools menu
Check Remote Execution Queue 9
Class Prediction 9
Clustering 8
Condition Tree 8
QT Clustering 9
Self-Organizing Map 9
Draw Expression Profile 9
Find Potential Regulatory Sequences 9
Find Significant Parameters 9
Find Similar Genes 9
Find Similar Samples 9
Gene Tree 8
K-Means 8
Principal Component Analysis 9
Script Editor 9
Statistical Analysis (ANOVA) 8
tree view 33
change color settings 37, 38
change features 37
change legend options 38
change thumbnails 32
create trees 30
display options 35
Eisen-like 31
magnify 33
make gene list 31
navigate subtrees 33
open 30
select subtree 31
view condition tree 36
view gene names 34
view gene tree 36
view nodes 34
view parameters 35
trust
described 61
set interpretation 62
Tukey 14
two-color experiments 36
two-way ANOVA
described 2
disproportional replication 22
equal replicates 23
equal replication 18
Friedman's test 22
interpret results 23, 25
no replicates 21, 22
perform 15
proportional replication 20
technical details 18

U
UniGene
database 4
update annotations 6
Unknown flag 27
Update Genome from Silicon Genetics window 29, 5
upregulated correlation 18, 14, 20

V
Venn diagram
color gene list 64
make gene list 18
save image 3
view
change 8
find gene 3
View External Program Details window 59
View menu
3D Scatter Plot 5
Animate 5
Animate Secondary 6
Array Layout 5
Bar Graph 5
Blocks 5
Compare Genes to Genes 5
Condition Scatter Plot 5
Display Options 6
Graph 5
Graph by Genes 5
Ordered List 5
Pathway 5
Physical Position 5
Remove Secondary Gene List 6
Scatter Plot 5
Show Average of Genes 6
Spreadsheet 5
Tree 5
Unsplit Window 5

Visible 6
Animation Controls 6
Hide All 6
Hide All in All Windows 6
Magnification 6
Navigator 6
Picture 6
Secondary Animation Controls 6
Secondary Picture 6
Show All 6
Show All in All Windows 6
Zoom Fully Out 5
Zoom In 5
Zoom Out 5
views
3D scatter plot 25
array layout 40
bar graph 12
blocks 7
compare genes to view 44
graph 9
graph by genes 46
ordered list 38
pathway 42
physical position 15
scatter plot 20
spreadsheet 50
tree 30

W

weighting coefficient 56
Westfall and Young 13
window
change color options 60
change Colorbar range 63
change default colors 67
change display options 57
change legend options 59
color by classification 66
color by gene list 67
color by parameter 65
color by secondary experiment 67
color by significance 64
color by Venn diagram 64
hide display elements 5
linked 1
no color 65
open 13
set expression 62
set trust interpretation 62
show display elements 5
solid color 66
split 1, 2
Window menu
Full Window 10
Tile Windows 10
wizard
Import Data 23
Import Data from Database 23
Import GeneSpring Zip 23
Import Genome 23

Z

zip file
export 6
import 5
zoom 9, 2